DOCUMENT RESUME

ED 385 588 TM 024 026

AUTHOR Schaeffer, Gary A.; And Others
TITLE Field Test of a Computer-Based GRE General Test. GRE Board Report No. 88-08P.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-93-07
PUB DATE Apr 93
NOTE 79p.; Cover title varies slightly.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Attitudes; *College Graduates; *Computer Assisted Testing; Equated Scores; Field Tests; Graduate Study; Higher Education; Item Response Theory; *Scores; Surveys; *Test Format; Test Items; Test Results
IDENTIFIERS *Graduate Record Examinations; Paper and Pencil Tests

ABSTRACT

This report contains results of a field test conducted to determine the relationship between a Graduate Records Examination (GRE) linear computer-based test (CBT) and a paper-and-pencil (P&P) test with the same items. Recent GRE examinees participated in the field test by taking either a CBT or the P&P test. Data from the field test indicated that examinees were able to navigate through the CBT with very little difficulty and that their overall reaction to the CBT was favorable. No substantive item-level mode effects were detected. No test-level mode effects were found for the verbal and analytical measures, but a small test-level effect was found for the quantitative measure. The raw-to-scale equating conversions of the CBT nearly matched those of the P&P version of the CBT for each measure, and therefore P&P conversions were used to report CBT scores. Six appendixes provide supplemental information, including item response theory parameter estimates, the plots of differences, the CBT questionnaire results, and seven additional tables. (Contains 31 tables, 13 figures, and 6 references.) (Author/SLD)

GRE®
RESEARCH

Field Test of a Computer-Based GRE General Test

Gary A. Schaeffer
Clyde M. Reese
Manfred Steffen
Robert L. McKinley
Craig N. Mills

April 1993

GRE Board Professional Report No. 88-08P
ETS Research Report 93-07

Educational Testing Service, Princeton, New Jersey

Field Test of a Computer-Based GRE General Test

Gary A. Schaeffer
Clyde M. Reese
Manfred Steffen
Robert L. McKinley
Craig N. Mills

GRE Board Report No. 88-08P

April 1993

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

Educational Testing Service, Princeton, N.J. 08541

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

Abstract

This report contains results of a field test conducted to determine the relationship between a GRE linear computer-based test (CBT) and a paper-and-pencil test (P&P) with the same items. Recent GRE examinees participated in the field test by taking either a CBT or a P&P test. Data from the field test indicated that examinees were able to navigate through the CBT with very little difficulty, and that their overall reaction to the CBT was favorable. No substantive item-level mode effects were detected. No test-level mode effects were found for the verbal and analytical measures, but a small test-level effect was found for the quantitative measure. The raw-to-scale equating conversions of the CBT nearly matched those of the P&P version of the CBT for each measure, and therefore P&P conversions were used to report CBT scores.

Acknowledgments

A number of individuals provided valuable expertise during this study. The authors thank the staff members of the Statistical Analysis area of the School and Higher Education Programs division at Educational Testing Service (ETS), in particular Louann Benton, Kathleen Haynie, Stefan Schuppan, and Marianne Vitella for their efforts in planning and performing analyses under severe time constraints. Marna Golub-Smith and Nancy Petersen provided very helpful technical and managerial input. Program directors Charlotte Kuh, Shelly Mendelowitz, Anne O'Meara, Susan Vitella, and Jayme Wheeler excellently coordinated the field test. Kathleen Carbery served as an outstanding GRE Systems contact. Charles Lewis and Rebecca Zwick generously provided top-notch statistical consultation (although any statistical shortcomings in the report are the responsibility of the authors). Other ETS staff members, too numerous to mention, provided thoughtful and careful reviews of drafts of this report.

Table of Contents

Table of Contents (continued)

1.0 Introduction

In June 1988, the Graduate Record Examinations (GRE) Board began consideration of a framework for research and development for a potential new Graduate Record Examination. The framework provided both short-term and long-term possibilities for a revised examination program. One set of possibilities concerned the development of a computerized adaptive General Test.

Computerized adaptive tests (CATs) offer a number of advantages to the program. These include shorter tests, immediate score reporting, and flexibility in scheduling administrations. The Board expressed interest in the development of such a test and, over the next year, funded a three-phase research and development project to achieve that goal. The third phase of the project was designed to provide a field test of a CAT. An operational CAT was to be introduced for limited administration if the field test results were supportive.

As is common in development activities, initial plans had to be revised. It was recognized that the development of a CAT involves two major changes in the presentation of a test. First, the mode of testing is changed. That is, instead of paper and pencil (P&P), a computer is used to present questions and record responses. Second, the testing paradigm is changed from a linear test where all examinees are administered the same set of items to an adaptive one where examinees are administered items that are tailored to their ability. The original proposal for a CAT field test was designed to assess the effects of these two changes simultaneously. As plans for the field test were being developed, however, it became apparent that each change individually has the potential to affect the relationship between the current General Test and the CAT. The original study design did not provide data that would allow determination of which factor is responsible for differences in performance (if differences were found). Therefore, the study was revised to provide a two-stage transition from P&P to CAT. The first step compares the linear P&P test to a linear computer-based test (CBT). The second step then compares the linear CBT to a CAT. The introduction of the CBT as an intermediate step in the transition was necessary to separate mode of testing and testing paradigm effects.

The revised plan called for the translation of an existing P&P test to computer format. This CBT is delivered on a computer, but is otherwise the same as the P&P version. This report summarizes findings of a field test conducted to determine the relationship of the CBT to a P&P test.

The field test consisted of a sample of volunteer GRE examinees taking a GRE General Test as a CBT. It also included a P&P group to control for retest effects. The study design allowed for examination of the effect of computer administration on examinee scores. CBT equating functions also were obtained so that CBT scores can be reported on the GRE scale when the CBT becomes operational. Data about timing and use of on-line tools (e.g., Review, Mark, Help) were collected. Reactions to the CBT also were obtained from field test examinees.

2.0 Methods

A sample of examinees from the October 1991 GRE national administration meeting certain criteria participated in the CBT field test. These volunteers were assigned to take either the CBT or another GRE P&P test several weeks following the national administration.

2.1 Data Collection Design

Forms T1GRE, T2GRE, and T3GRE were spiraled at the October 1991 national administration.[1] Spiraling is designed to assure that randomly equivalent groups take each form. From the examinees taking Form T2GRE at this administration, volunteers were recruited to participate in the field test. The national administration test form spiral was established such that approximately 60% of examinees took T2GRE and about 20% each took T1GRE and T3GRE. This maximized the number of potential field test examinees, while still allowing enough examinees to take the other two forms to permit operational equating and differential item functioning (DIF) analyses.

As part of the field test, examinees took another General Test between one and nine weeks after the October 1991 national administration: either a CBT version of T1GRE (CBT) or a field test P&P version of T1GRE (S-T1GRE). Both the CBT and the S-T1GRE field test administrations occurred on average about 5.5 weeks after the national administration. However, the CBT administrations were spread throughout the 9-week interval, but the S-T1GRE administrations occurred within about a 2-week period.

The CBT, T1GRE, and S-T1GRE have the same operational sections. The operational sections of the CBT and S-T1GRE were in the same order, but the order of sections in the field test forms differed slightly from the section order of T1GRE. In T1GRE the nonoperational section appeared as the fifth section, but in the field test forms the six operational sections were in the first six slots. In addition, the nonoperational section of the CBT appeared last, and S-T1GRE did not have a nonoperational section. Each form had 38 items in each verbal section, 30 items in each quantitative section, and 25 items in each analytical section. The orders of sections are listed below.

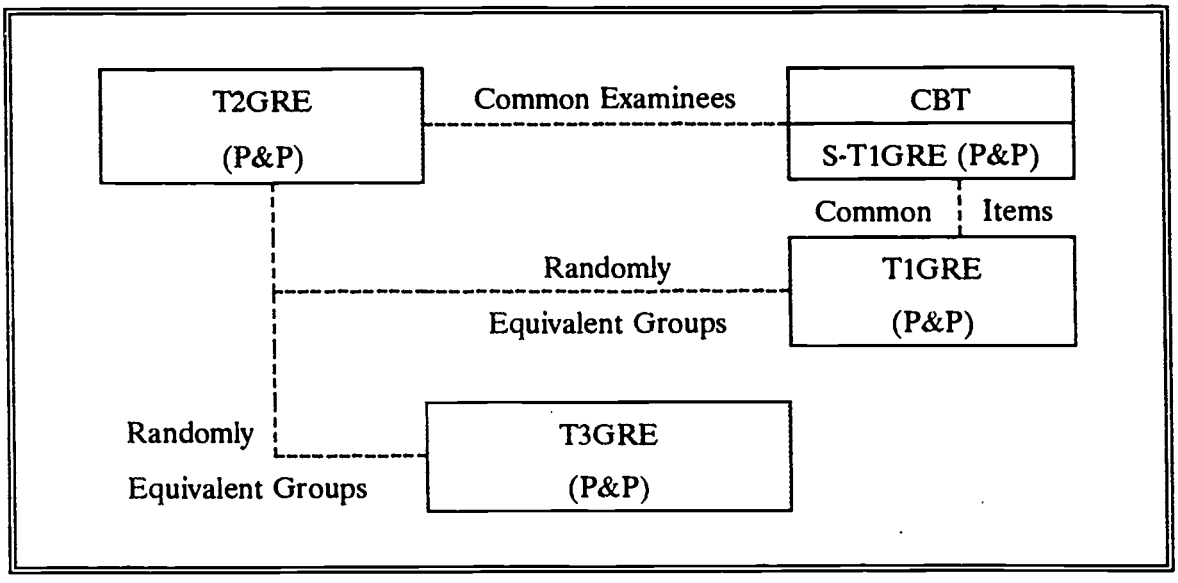
Section	CBT	S-T1GRE	T1GRE
1	Quant1	Quant1	Quant1
2	Verbal1	Verbal1	Verbal1
3	Analytical1	Analytical1	Analytical1
4	Quant2	Quant2	Quant2
5	Verbal2	Verbal2	Non-oper
6	Analytical2	Analytical2	Analytical2
7	Nonoper	---------	Verbal2

Figure 2.1 illustrates the data collection design for the field test.

[1] These are not the operational test form designations.

Figure 2.1

Field Test Data Collection Design



NOTE: Shaded boxes represent portions of the regular October 1991 administration of the GRE General Test. Unshaded boxes represent portions of the field test.

CBT field test examinees took the CBT several weeks after they took an operational P&P GRE. Performance on the CBT, therefore, may have been influenced by a retest effect (e.g., practice). In fact, this retest effect would be confounded with any effect due to the computer mode of the test administration. To help isolate and quantify any retest effect, the field test P&P component (S-T1GRE) was included in the design. Differences in scores for examinees who took P&P tests both operationally and at the field test would be due to a retest effect. The retest effects component was used to assist in interpreting item-level and test-level mode effects. However, a general limitation of the data collection design was that for analyses that directly compare CBT to T1GRE, any effects due to mode and/or retest cannot be completely disentangled.

2.2 Field Test Centers

The main concerns when selecting centers to administer the field test were (a) obtaining an adequate number of examinees for the analyses and (b) obtaining a sample that approximately reflected the GRE test-taking population in terms of, at least, ability, gender, ethnic group ratios, and geographic dispersion. In addition, centers were selected that historically had included relatively large numbers of Black and Hispanic GRE examinees. Based on past volumes, demographic characteristics, and score distributions, test centers at the following locations were included in the field test:

Atlanta, GA	Emeryville, CA	Lawrenceville, NJ	Pasadena, CA
Austin, TX	Evanston, IL	Los Angeles, CA	Tempe, AZ
Baton Rouge, LA	Greensboro, NC	Miami, FL	Trenton, NJ
Cambridge, MA	Houston, TX	Norfolk, VA	Washington, DC

All test centers were located near college or university campuses to facilitate the recruitment of examinees. The CBT was administered at all of these locations except Los Angeles and Trenton.

Because only 200 examinees were recruited to take S-T1GRE, (as opposed to more than 1,000 for the CBT), for economic and administrative reasons it was decided not to administer S-T1GRE at all CBT test centers. Instead, S-T1GRE was administered only in Houston, Trenton, and Los Angeles. Centers at these locations were selected because they were geographically diverse and, based on previous GRE data, were representative of the groups testing in the areas where CBT centers were located in terms of mean GRE scores and also gender, Black, and Hispanic examinee proportions.

2.3 Examinee Recruitment

At the October 1991 administration, examinees at testing centers within 75 to 100 miles of the centers listed above were given written invitations to participate in a GRE study in which they would take another GRE test in either CBT or P&P format. They were told they would be randomly assigned to test format. They also were informed that only examinees who had taken T2GRE at the national administration would be considered for inclusion in the study. Incentives included $50 and the option of reporting their field test scores. The original recruitment goal was to obtain at least 1,000 CBT test takers and about 200 examinees to take S-T1GRE.

Examinees who indicated that they were interested in participating in the field test were classified as to whether they met specified analysis criteria, based on their October 1991 data already on the GRE files. If they did, they were assigned a field test administration day and time that was convenient for them. If they did not meet the criteria, they were not accepted into the study (except during the first week or so of the field test, when it could not be determined if volunteers met the criteria; at that time, all volunteers were accepted into the study, although those not meeting the criteria were not included in the analyses). The criteria for acceptance into the field test were based on examinees' responses to the background questions they had completed for the October 1991 national administration. The following criteria are required for inclusion into GRE operational equating samples and were applied here since a major goal of the field test was to equate the CBT:

1. test at a regular national administration
2. are U.S. citizens
3. consider English to be their best language
4. test at a domestic test center
5. register for the October 1991 regular administration through the regular registration process
6. have an appropriate irregularity code
7. mark as a reason for taking the GRE General Test at least one of the following:
   a. admission to graduate school
   b. fellowship application requirement, or
   c. graduate department requirement

The original field test plan included oversampling Black and Hispanic examinees to participate in the field test. This would have permitted more extensive field test analyses within each of these subgroups. However, due to the difficulty in recruiting a representative sample of sufficient size for the primary analyses, the short time line to complete the field test, and the small number of minority examinees in the pool, efforts to oversample Black and Hispanic examinees sufficiently for extensive within-group analyses were not successful. On the other hand, a sample representative of the proportions of Black and Hispanic examinees typically taking the GRE was included in the study, and descriptive statistics based on these groups are presented.

2.4 Description of the CBT

The CBT form used in this study was paper-and-pencil T1GRE converted for computer delivery. This particular form was selected for conversion to a CBT primarily because it (a) had already been administered at past national administrations and was shown to be typical of other GRE forms in terms of item and test statistics and (b) had not been disclosed at earlier administrations. For the purpose of assessing comparability of CBT and P&P, the CBT items were made to look like the P&P T1GRE items as much as possible. These items were transcribed onto the computer following ETS style guidelines and were carefully proofed.

After several tryouts involving ETS staff and a small sample of local college students, the CBT was deemed ready for the field test. Electronic copies of the CBT, with accompanying software, were sent to the field test test centers. Each test center was equipped with between 4-10 computers. The center administrator used a combination of manual and computer procedures to control operations and deliver tests at the center, including installing item and test information, signing examinees onto computers, and ending the testing session.

Once examinees had provided sufficient identification at the test center, the center administrator allowed them to begin the test. Examinees used a mouse to navigate on the computer and to record responses. Four tutorials were presented on the computer to all examinees before the operational items were presented. Examinees could determine how much time they wanted to spend on each tutorial (however, once they left a tutorial they could not return, although tutorial information was available in the Help tool). These included tutorials on the mouse, testing tools, selecting an answer, and scrolling. Eight testing tools were available to CBT examinees throughout the test, each with its own icon[2]. The tools allowed examinees to do the following:

Quit:	quit the test
Exit:	exit the section
Time:	show/hide time remaining in section
Review:	go to any item in section/check status of items in section
Mark:	mark an item for later review
Help:	view previously presented information, i.e., directions, summary of tutorials
Prev:	view screen previously seen
Next:	move to next screen

[2] During the seventh (i.e., nonoperational) section, the Review tool was turned off.

Examinees were given 32 minutes to complete each section.[3] The amount of time examinees spent on the tutorials is shown in Appendix A. Directions screens were used to display test, section, item group, and item set directions. An optional 10-minute break was offered to examinees between sections 3 and 4 of the CBT. At the end of the CBT session, examinees were given a questionnaire to complete. The total amount of time spent by CBT examinees on orientation activities, administration of the test, and completion of the questionnaire was usually between 4.5 and 5 hours.

2.5 Field Test Questionnaire

Upon completion of the CBT, examinees were asked to complete a questionnaire on paper. The questionnaire asked about general computer experience, the specific CBT test-taking experience, and test-taking strategies and preferences. Questionnaire results are presented in Section 9.0.

3.0 Description of Samples

A primary purpose of the field test sampling scheme was to obtain samples of CBT and S-T1GRE examinees that were similar to each other and to the October 1991 administration examinees in terms of background and score characteristics. Obtaining reasonably representative field test samples enhances the generalizability of the CBT findings to the GRE population.

3.1 Field Test Sample Background Information

Of the total of 1,153 examinees who took the CBT, 1,017 met the selection criteria. Of the 196 examinees who took Form S-T1GRE, 184 met the selection criteria. Table 3.1 shows the gender and ethnicity composition of the CBT, S-T1GRE, and October 1991 administration samples[4] that met the criteria.

Table 3.1
Gender and Ethnicity Proportions

	Female	Male	Asian	Black	Hispanic	White
CBT Sample	61%	39%	6%	8%	5%	77%
S-T1GRE Sample	55%	45%	10%	8%	9%	71%
Oct. 91 Admin.	61%	38%	4%	5%	4%	85%

[3] The P&P tests allow examinees 30 minutes per section. For the CBT, a total of 32 minutes per section was allowed to account for the time it took to refresh computer screens between items (generally less than 2 seconds per item).

[4] Because the three operational forms were spiraled in October 1991, examinees who took T1GRE are representative of examinees who took all three forms during the operational administration; therefore, background data from T1GRE examinees are presented in Table 3.1 as October 1991 administration data.

The gender proportions were similar in the field test and October 1991 samples, and the field test samples had slightly greater proportions of ethnic minority examinees than did the October 1991 GRE national population.

Another recruitment goal was that the field test samples be similar to the October 1991 administration sample in terms of ability as defined by GRE test scores. Table 3.2 presents means and standard deviations of T2GRE operational number-right scores for the CBT sample, the S-T1GRE sample, and the October 1991 administration sample who took that form. These data suggest that the two field test samples were similar to each other. For both field test samples, the mean GRE scores on each measure were slightly higher than those for the October 1991 administration. The differences between the field test samples and the national administration were approximately a quarter of a standard deviation.

Table 3.2
National Administration Score Distributions

	Mean (and S.D.) of October 1991 GRE Number-Right Scores--T2GRE		
	Verbal	Quantitative	Analytical
CBT Sample	49.6 (12.4)	36.1 (10.9)	32.1 (8.5)
S-T1GRE Sample	48.9 (12.6)	36.7 (11.8)	31.2 (9.3)
Oct. 1991 Admin	46.4 (12.3)	33.2 (11.1)	30.0 (8.6)

3.2 Field Test Score Distributions

For the CBT and S-T1GRE field test samples, a summary of number-right field test scores is presented in Table 3.3. The data for the two samples are similar and suggest a reasonable level of achievement and spread in scores. Note that number-right scores earned at the October 1991 administration on T2GRE cannot be compared to these scores because the scores are from different forms.

Table 3.3
Field Test Score Distributions

	Number	Mean (and S.D.) of Field Test Number-Right Scores		
		Verbal	Quantitative	Analytical
CBT Sample	1,017	49.9 (12.0)	41.7 (10.6)	33.8 (8.4)
S-T1GRE Sample	184	49.9 (12.2)	43.4 (10.7)	32.5 (9.4)

4.0 IRT Calibrations

4.1 T1GRE Calibration and Scaling

Form T1GRE, administered as part of the October 1991 regular GRE General Test administration, was calibrated using the three-parameter logistic (3PL) model, given by

$$P_i(\theta_j) = c_i + (1 - c_i)/(1 + \exp(-1.702 a_i(\theta_j - b_i))) , \quad (1)$$

where a_i is the item discrimination parameter for item i, b_i is the item difficulty parameter for item i, c_i is the item lower asymptote, or pseudo-guessing, parameter for item i, and θ_j is the ability parameter for examinee j.

The first step in the calibration process was to score the items. This was accomplished by assigning a score of 0 to all incorrect answers, 1 to all correct answers, 2 to all omitted items, and 3 to all items not reached. For the purpose of this calibration, omits were defined as unanswered items that were followed by answered items. Not-reached items were those items not answered and not followed by answered items.

These scored items were then calibrated using the LOGIST 6 (Wingersky, Patrick, & Lord, 1988) estimation program. The two sections of each measure (verbal, quantitative, and analytical) were calibrated together, and the measures were calibrated separately.

Finally, the calibration of T1GRE was placed on the GRE base IRT scale via common item scaling. The item characteristic curve method (Stocking & Lord, 1983) of common item scaling was used to scale this calibration to the base GRE form. The item characteristic curve method scaling places two calibrations on the same scale by minimizing the differences between the test characteristic curves for the items in common to the two calibrations.

4.2 CBT Calibration and Scaling

The CBT, the computer-administered version of T1GRE, was also calibrated using the 3PL model. Item scoring was performed in a slightly different manner. As before, all incorrect answers were assigned a score of 0, all correct answers were assigned a score of 1, all omits were assigned a score of 2, and all not reached items were assigned a score of 3. However, omits were not defined here as unanswered items followed by answered items. Because the items were administered on the computer, it was possible to determine whether the examinee actually viewed each item. Therefore, omits were defined as items viewed but not answered. Not-reached items were those items not viewed by the examinee.

As with the P&P form, the CBT form was calibrated using the LOGIST 6 estimation program. Sections for a measure were calibrated together, but measures were calibrated separately. Calibrations again were placed on the base GRE IRT scale using the item characteristic curve method. Lists of CBT item parameter estimates and plots of CBT and T1GRE item parameter estimates are in Appendix B.

4.3 Criterion Calibration and Scaling

A direct comparison of the item parameters for T1GRE obtained under CBT and P&P conditions, in the absence of other information, would be difficult to interpret. Differences in item parameters might be due to variations in the size or representativeness of the calibration sample or mode effects. In order to provide a baseline for comparison where no mode effects were present, a sample from a P&P administration of approximately equal size to that used for the CBT calibrations was used to independently calibrate and scale form T1GRE. Note that Form S-T1GRE was not calibrated because of the small sample size.

The criterion calibration process was very simlar to that just described for the October 1991 regular GRE General Test data with two exceptions. First, the calibration sample consisted of a random sample of 1,045 examinees selected from the full October 1991 administration calibration sample of 13,594 examinees. Second, the calibrations derived from the 1,045 examiness were scaled directly to the scaled calibrations derived from the full October 1991 sample rather than to those based on the base form calibrations.

Thus, if the item parameters obtained from the CBT sample and the criterion calibration samples differ from the base form parameters in a similar fashion, it can be inferred that CBT based differences are consistent within sampling variation and that negligible mode effects are present. The degree to which the CBT-based parameter differences are appreciably larger than the criterion sample differences would be evidence of mode effects.

5.0 Item-Level Mode Effects

Conceptually, mode effects are defined as differences in examinee performance produced by the mode of presentation (CBT or P&P) of an item or test. For example, for groups of examinees with similar abilities, if an item is easier in one mode than the other, this is evidence of a mode effect for that item. A test-level mode effect (see section 6.0) might be evidenced by examinees producing lower scores on the CBT than they would have produced on a P&P edition of the same test. Evidence of item-level mode effects is not prerequisite to presence of test-level mode effects. The assessment of item-level mode effects was performed using IRT, classical, Mantel-Haenszel, and logistic regression statistics.

5.1 Item-Level Mode Effects--IRT

The search for items displaying mode effects using IRT used an item flagging process based on statistical analyses performed on the individual item parameters and on the item characteristic curves (ICCs). The analyses focused on comparing estimated item parameters and ICCs from the CBT to those obtained for T1GRE in an attempt to determine what items were functioning differently in the two modes. The analyses of item parameter estimates were designed to assess changes in the characteristics of items, such as discrimination and difficulty, while analyses of ICCs were designed to determine the effect of changes in item characteristics on examinees at different ability levels. The analyses performed on the CBT calibration were repeated for the criterion calibration to provide a baseline for evaluating the magnitude of the differences. Note that since the CBT item parameters were put on scale through the T1GRE item parameters, any main effect due to mode (that is, an effect that affects all items similarly) would be masked in the scaling process. Therefore, only item-by-mode interactions can be detected using this procedure.

Items were flagged as appreciably different if the difference between b-values was larger than the standard deviation of the b-value differences--this does not constitute a test of significance of the difference, but serves as a convenient mechanism for determining whether relatively large changes occurred. In addition to the statistics computed on the item parameter estimates and ICCs, overlay item characteristic curve plots (sometimes called item-ability regression, or IAR, plots) were constructed and examined. The IAR plots were constructed using the procedure described by Kingston and Dorans (1985), with one modification. The IAR plots for the two calibrations of the same item were overlaid on the same set of axes, thus providing a direct visual comparison of the two ICCs.

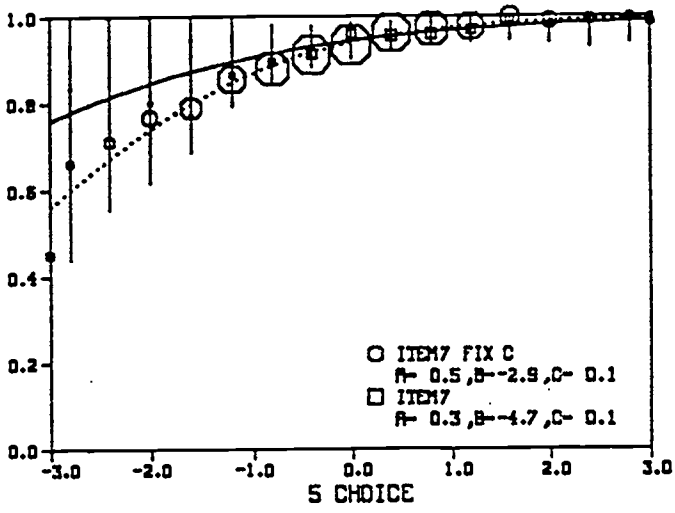
Although a number of items in each measure were flagged for closer inspection, further analysis in each case indicated only differences in item functioning that would be expected from normal sample fluctuation. Flagged items did not appear to be of any particular item type or content, nor was there anything systematic about their location in the test. Rather, the flagging appeared to result from factors such as instability in the item b- and c-parameter estimates due to small sample sizes at extreme ability levels.

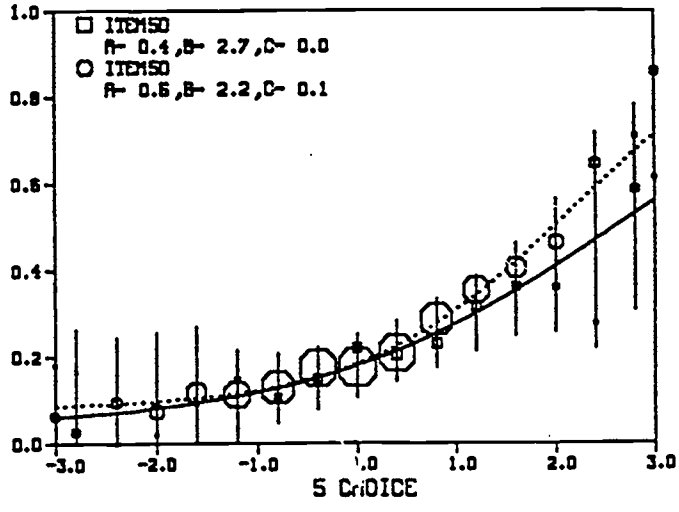
For the analytical measure, for example, five items appeared appreciably more difficult on the computer than in the P&P format, and two items appeared appreciably more difficult in the P&P mode. But, in almost every case the changes in difficulty (whether increase or decrease) could easily represent statistical artifacts of the estimation process. This is illustrated in Figure 5.1, which shows the item characteristic curves for four analytical items that typify the results for all measures. For each item shown, the solid line shows the ICC for the CBT calibration, and the dotted line shows the ICC for the T1GRE calibration. Vertical lines represent the standard error of estimate. The hexagons and squares are centered on the corresponding observed proportion correct, and their sizes are a function of sample size.

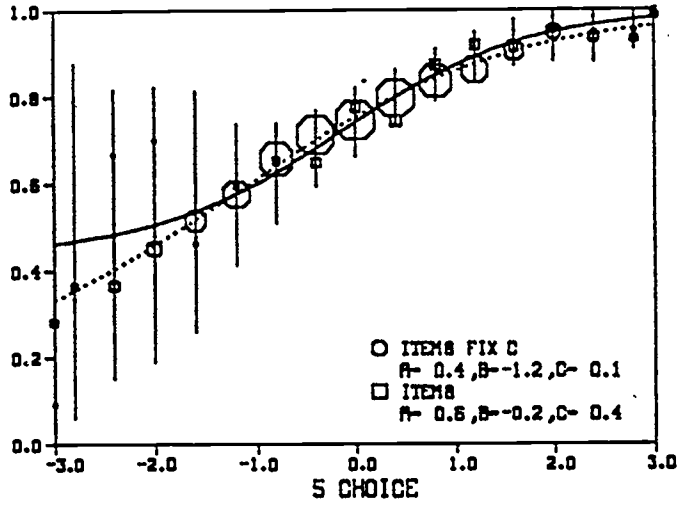
The ICCs shown for item 7 typify the case of a very easy item having an unstable difficulty estimate. Sample sizes at low ability levels tend to be small, and the standard error of estimate for item difficulty values in this range tend to be rather large. In the case of item 7, the b-value for the CBT calibration decreased by an amount that is within the range of estimation error resulting from the smaller calibration sample size. Thus, the item appears to be easier on the computer than in the P&P format. The ICCs for item 50 show a case where the b-value of a very hard item increased by an amount that is within the range of the increased estimation error.

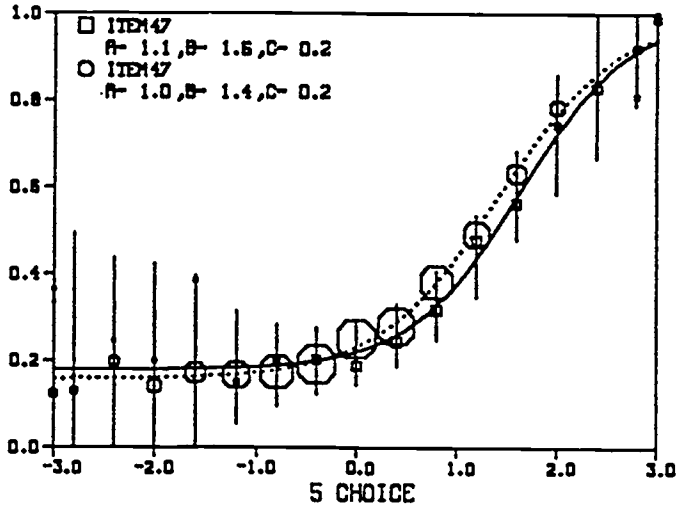
The ICCs for item 8 show how errors in the estimation of the c-parameter can result in changes in the b-value. In this case, the b-value increased for the CBT mode as a result of an increase in the c-parameter estimate (this was confirmed by recalibrating the CBT data holding the c-values fixed at the T1GRE calibration values, resulting in elimination of the b-value differences). It is important to note that for situations such as those presented here, the ICCs for the two calibrations are very similar throughout most of the ability range, despite the rather noticeable change in b-values. This indicates that the change in b-values does not indicate a major change in the way the items are functioning.

Figure 5.1
Item Characteristic Curves for Four Analytical Items









The ICCs shown for item 47 indicate there was very little change in the item parameter estimates across the two calibrations. The majority of items for all three measures showed a pattern similar to item 47.

For the quantitative measure, 11 items were appreciably more difficult on the computer, and 4 items were appreciably harder in P&P mode. For the verbal measure, 9 items appeared to be appreciably harder when administered on the computer than when administered via P&P, and 16 items appeared to be more difficult when administered via P&P. For both measures, the patterns of ICCs closely followed those found for the Analytical measure. Also, the MIMD statistic, an indication of overall bias, indicated no bias in any of the three measures (see Appendix C). This is an indication that the scaling process was effective.

Table C.1 in Appendix C summarizes all of the statistics computed to compare the CBT and criterion calibrations to the T1GRE calibration. Results are presented for each measure for each item parameter and the ICCs. Statistics computed for the criterion sample represent the degree to which differences are expected to be nonzero as a result of sampling variation. For all measures and item parameters the mean deviation (MD) and mean absolute deviation (MAD) are of the same order of magnitude for the CBT and criterion calibrations. Several of the maximum deviations (Max D+, Max D-) are larger for the criterion calibrations. Taken as a whole, there is little evidence to support the presence of mode effects in the CBT calibrations.

Based on the results of the analyses performed on the IRT calibrations, the following conclusions seem warranted. First, the differences in item parameter estimates and ICCs for the two calibrations of this form are small and do not appear to be systematic, indicating that there is no need to remove any items from the common item block for IRT scaling purposes. Most differences of appreciable size can be explained as a result of the instability of item parameter estimates.

Second, although the CBT and T1GRE calibrations produced very similar results, the differences in item parameter estimates and ICC's for the CBT calibrations generally were larger than those based on the criterion calibration. These differences are to be expected to some degree because the criterion calibration group was a random sample of the T1GRE group but the CBT group was not. Also, all examinees in the CBT sample were repeaters (i.e., had taken the GRE previously), but only a small fraction of the criterion calibration sample was repeaters. Nevertheless, it is possible that there was something different about the CBT calibrations that resulted in these differences. Although the differences found did not indicate the need for recalibrating items using CBT data for this particular form, they do indicate a need for additional research to determine whether recalibration might be warranted for future CBT forms.

5.2 Item-Level Mode Effects--Classical Statistics

Item difficulty and discrimination were examined using deltas and biserial correlations. Delta is an index of item difficulty that is a transformation of the percentage of the group that answered the item correctly; delta values increase with increasing item difficulty. Equated deltas are deltas adjusted for the ability level of the group that took the item; therefore, items with similar equated deltas are considered to be similar in difficulty, regardless of the ability of the groups that took the items. Equated deltas range from about 6 (very easy) to about 20 (very difficult).

The equating transformation of CBT and S-T1GRE deltas resulted in the same mean and standard deviation for each form (which were the same as for T1GRE). These values are listed in Table 5.1. Note, however, that the delta equating process does not necessitate that each individual item have the same equated delta across forms.

Table 5.1
Equated Delta Distributions

	Equated Delta	
	Mean	S. D.
Verbal	11.9	2.5
Quantitative	11.6	2.3
Analytical	12.5	2.5

The biserial correlation for an item is an approximation of the product-moment correlation between the total score of which the item is part and a theoretical normally distributed continuous underlying variable hypothesized to have determined the right or wrong response to the item. Items selected for inclusion in the General Test have biserial correlations based on pretest statistics that are always positive. Table 5.2 lists the mean biserial correlations by measure for the three tests: CBT, S-T1GRE, and T1GRE. This table indicates that the mean biserial correlations on the verbal and analytical measures were slightly higher for both field test samples than for the T1GRE sample.

Table 5.2
Mean (and S.D.) of Biserial Correlations

	CBT	S-T1GRE	T1GRE
Verbal	.53 (.10)	.52 (.12)	.49 (.09)
Quantitative	.58 (.10)	.58 (.14)	.57 (.10)
Analytical	.56 (.08)	.59 (.11)	.52 (.08)

Form T1GRE equated deltas and biserial correlations were compared to CBT equated deltas and biserial correlations. (Form S-T1GRE equated deltas and biserials correlations were not included in this comparision because of the small sample size.) Note that this comparison includes any retest effects as well as mode effects. Table 5.3 shows the means and standard deviations of the CBT and T1GRE differences in equated deltas and biserial correlations for each measure and section. The mean differences in equated deltas by section were largest for the analytical measure; however, they were only about a tenth of a delta point. For the verbal and analytical sections, the biserial correlations for the CBT were on average .03-.05 higher than for T1GRE.

Table 5.3
Distributions of Differences in Equated Deltas and Biserial Correlations

CBT Section	Number of Items in Comparisons*	Mean (and S.D.) of Differences in Equated Deltas (CBT- T1GRE)	Mean (and S.D.) of Differences in Biserial Correlations (CBT - T1GRE)
Verbal Sec. 2	33	.02 (.38)	.04 (.05)
Verbal Sec. 5	34	-.02 (.27)	.03 (.06)
Quant. Sec. 1	30	.05 (.45)	.00 (.05)
Quant. Sec. 4	29	-.05 (.37)	.02 (.04)
Analyt. Sec. 3	24	-.11 (.53)	.03 (.05)
Analyt. Sec. 6	25	.10 (.43)	.05 (.04)

* Number of items in comparisons includes only items where delta values and biserial correlations were calculated; these statistics were not calculated for items for which more than 95% of the sample marked the correct answer, because of the resulting instability of the statistics.

It should also be noted that the largest equated delta differences on any item were about one delta point for each measure. More CBT biserial correlations were larger than their T1GRE counterparts than vice versa, and the vast majority of the differences in biserial correlations were less than 0.10. In fact, for items with lower CBT than T1GRE biserial correlations, the greatest difference was only 0.13.

These data suggest that there were only slight shifts in item difficulty and item discrimination in going from P&P to CBT mode, and that these shifts probably are of no practical significance in terms of CBT mode effects.

5.3 Item-Level Mode Effects--Mantel-Haenszel

Mantel-Haenszel (MH) procedures are designed to examine differential item functioning while controlling for ability (Holland and Thayer, 1988). Usually, MH is used to compare gender or ethnic group performance on a set of items (see Section 10.3 for MH gender analyses of the CBT). In the present case, MH is used to compare CBT and P&P performance on an item with examinee ability defined by October 1991 GRE scores. MH procedures have been shown to be the most powerful for detecting uniform DIF, that is, when differences between groups are the same at all ability levels. However, if nonuniform mode effects are present (i.e., when there is an interaction of ability and group membership), MH may underestimate the degree of difference between the two testing modes. To guard against this, a logistic regression procedure also was employed to assess DIF (see Section 5.4).

For the MH analyses, the ETS criteria typically used for flagging items exhibiting DIF were employed. "C" items show more DIF than "B" items, which show more DIF than "A" items. For an item to be flagged as a C item, the absolute value of its DIF statistic (labeled "MH D-DIF" at ETS) must be at least 1.5 and also must be significantly greater than 1.0 at the .05 significance level. On operational programs, only C items require action before score reporting; C items will be reported here.

The field test design included three groups of examinees taking T1GRE items: the CBT group, the S-T1GRE group, and the October 1991 T1GRE group. Both the CBT, S-T1GRE and the CBT, T1GRE comparisons were considered. However, the CBT, T1GRE comparison confounded mode and retest effects. Only the CBT, S-T1GRE comparison directly addressed mode effects. The results of these comparisons are presented below.

In the MH analyses of the CBT and S-T1GRE groups, one verbal item was flagged (favoring the CBT group), one quantitative item was flagged (favoring the S-T1GRE group), and one analytical item was flagged (favoring the CBT group). This number of items flagged was not above chance level, and no explanation in terms of item types or item presentation were found regarding the items flagged. In the CBT, T1GRE comparison, where mode and retest effects are confounded, no items suggesting disfavor toward the CBT group were flagged, and six items were flagged favoring the CBT group. It was concluded that no substantive item-level mode effects were detected using the MH method.

5.4 Item-Level Mode Effects--Logistic Regression

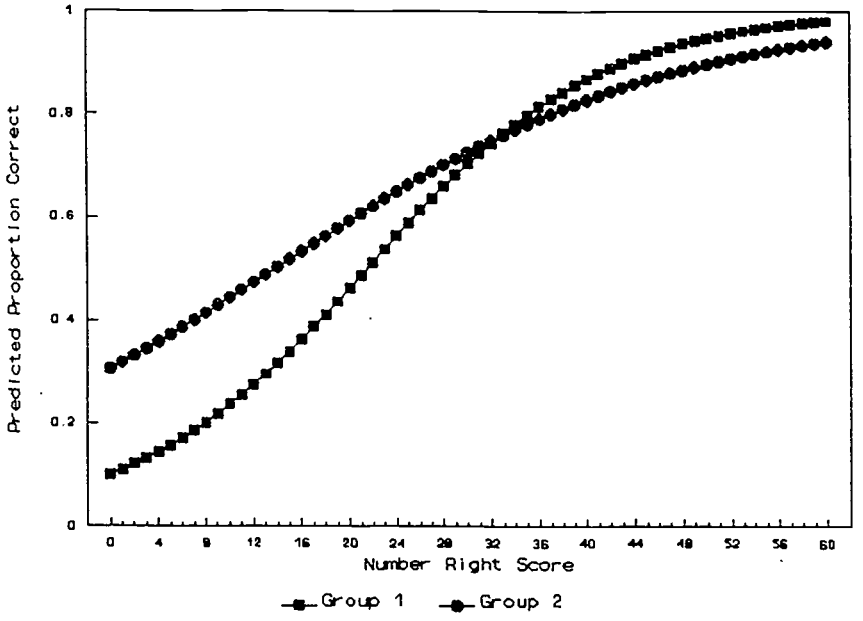
A logistic regression (LR) model also was employed as a DIF method to compare CBT and P&P performance (Swaminathan & Rogers, 1990). These LR analyses tested a regression model to predict performance on an item based on CBT or P&P group membership, ability defined by October 1991 GRE scores, the interaction between ability and group, and an intercept term. LR is particularly well suited for uncovering non-uniform mode effects, that is, when there is an interaction of ability and group membership. Unlike for classical item statistics, statistical significance tests are available. Regression parameters greater than zero at the .01 level were flagged.

Several possible patterns of effects are possible. A significant group effect and no significant interaction may be found where one group has a higher predicted proportion correct than the second group. Statistically significant group-by-score interactions imply that the differences between groups in predicted proportion correct varies across the score range. Figure 5.2 illustrates an example of a graph of a hypothetical item of a 60-item test with statistically significant group-by-score interaction terms. Note that Group 2 is favored for examinees with very low scores, and that at the upper end of the scale Group 1 is slightly favored. For examinees falling into the middle of the score range, the differences between the two groups are small.

Items were flagged that had statistically significant group or group by score interaction terms. For items with both significant group and interaction terms, only the interaction term is reported. For the verbal measure using the CBT/S-T1GRE comparision, three items were flagged, one favoring the CBT (i.e., a significant group effect), and two interactions. The CBT/T1GRE comparison also was considered, even though potential retest and mode effects were involved. As expected, more items were flagged: 1 item favored the CBT, and 12 items were flagged with significant interactions. The two interaction items flagged in the CBT/S-T1GRE comparison were not flagged in the CBT/T1GRE comparison.

The relatively small sample size of the S-T1GRE group may have affected the power of the significance tests performed in the CBT/S-T1GRE comparison. As such, items that were flagged in the CBT/T1GRE comparison were examined in the CBT/S-T1GRE comparison with a larger significance level. Only 3 of the 13 items flagged in the CBT/T1GRE comparison were flagged in

Figure 5.2
Logistic Regression Plot for a Hypothetical Item



the CBT/S-T1GRE comparison using a significance level of .10. No patterns of item type or item presentation were detected for any of the groups of flagged items.

For the quantitative measure using the CBT/S-T1GRE comparison, one item was flagged that favored the S-T1GRE group. Ten other items were flagged in the CBT/T1GRE comparison; nine had significant interactions, and one favored the CBT group overall. Of the 10 items, only one was significant at the .10 level in the CBT/S-T1GRE comparison.

For the analytical measure using the CBT/S-T1GRE comparison, no items were flagged. In the CBT/T1GRE comparison, 14 items were flagged; 6 interactions and 8 favoring the CBT overall. Only two of these items were flagged at the .10 level in the CBT/S-T1GRE comparison, and they could not be explained in terms of item type or item presentation.

Results also were obtained from the S-T1GRE/T1GRE comparison, which assess retest effects only. Three verbal, two quantitative, and two analytical items were flagged. The significant group effects favored the S-T1GRE group. The pattern of results matches almost all of the results found in the CBT/T1GRE comparison for each measure, suggesting again that the items flagged in the CBT/T1GRE comparison may be due more to retest than to mode.

The primary CBT/S-T1GRE logistic regression analysis of each measure flagged items at chance rate and no patterns were detected in the items that were flagged in terms of item types or item presentation. Analyses of the CBT/T1GRE and S-T1GRE/T1GRE comparisions yielded no evidence of substantive item-level mode effects. Rather, the effects found may be due mostly to

retest effects. As with the MH analyses, the LR procedure uncovered no practically significant item-level mode effects.

5.5 Conclusions

Based on the cumulative results of these analyses, several conclusions seem warranted. First, although differences in item performance were found by all methods used, the magnitudes of the differences were small with respect to expected variations. Most differences of appreciable size can be explained as a result of the instability of item parameter estimates. Second, items were not flagged consistently by multiple methods, as can be seen in Appendix D, which contains a list of all items flagged by the logistic regression, Mantel-Haenszel, and IRT methods. Finally, examination of the mean differences suggests that small differences are canceled out across items.

6.0 Test-Level Mode Effects

Test-level mode effects refer to whether examinees of comparable ability perform differently in one mode or the other on the test as a whole. Test-level mode effects was assessed by comparing examinee performances on the CBT with the two P&P versions, T1GRE (sometimes referred to as T1) and S-T1GRE (sometimes referred to as S-T1).

6.1 Score Distributions

Figures 6.1 - 6.3 show the percentage of examinees who scored below each number-right score on each test and measure. The shapes of the curves of each sample are similar. However, the curves for the T1GRE group are slightly above the curves for the other two groups, indicating that scores on T1GRE were somewhat lower. This indicates that the two field test samples were slightly more able than the operational sample, as was discussed in Section 3, and that the field test samples may have benefited from retest effects.

Figure 6.1
Percent Below Distributions- Verbal

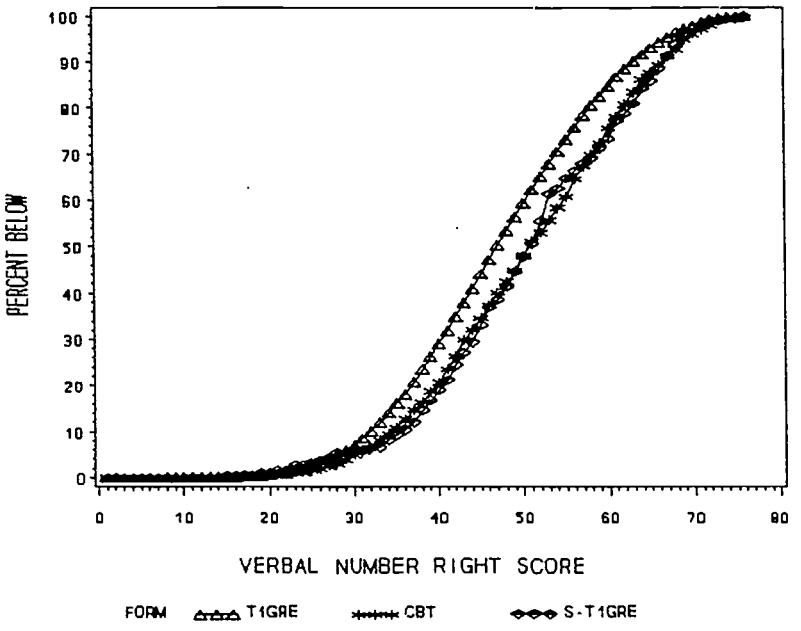


Figure 6.2
Percent Below Distributions- Quantitative

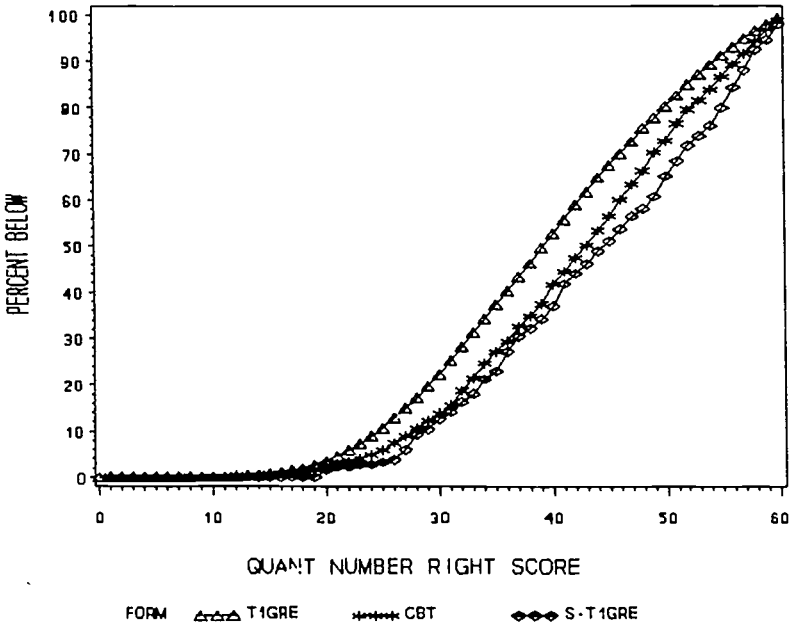
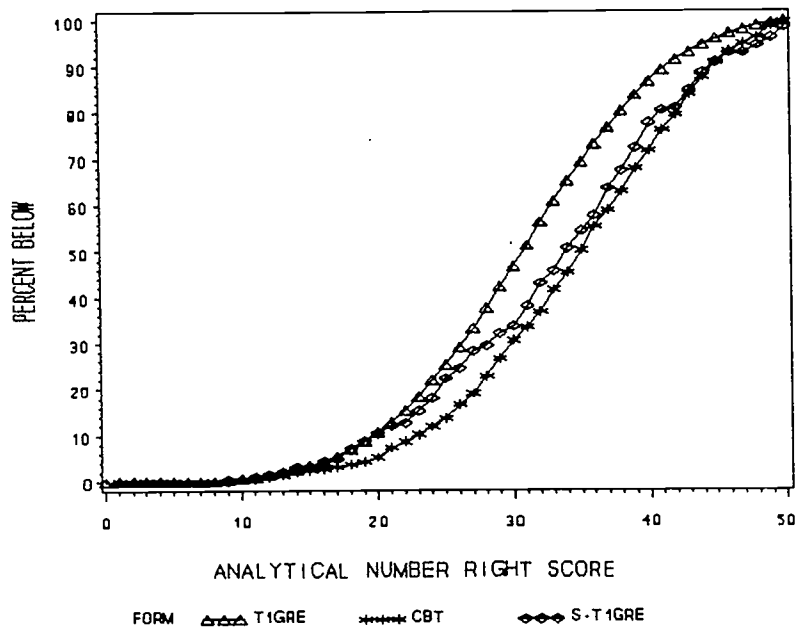


Figure 6.3
Percent Below Distributions- Analytical



6.2 Test Section Summary Statistics and Intercorrelations

Table 6.1 presents means and standard deviations of section scores for each measure for the CBT, S-T1GRE, and T1GRE samples. For each sample, the means of the first quantitative and analytical sections are higher than those of the second sections of those measures; the reverse is true for each sample for the verbal measure.

Table 6.1
Section Score Means (and Standard Deviations)

Field Test Section Number	CBT Sample	S-T1GRE Sample	T1GRE Sample
Verbal Sec 2	24.6 (6.5)	24.6 (6.7)	23.1 (6.4)
Verbal Sec 5	25.3 (6.0)	25.2 (5.9)	23.5 (5.9)
Quantitative Sec 1	20.9 (5.5)	22.0 (5.3)	19.5 (5.7)
Quantitative Sec 4	20.8 (5.5)	21.5 (5.6)	19.2 (5.7)
Analytical Sec 3	17.3 (4.5)	16.6 (4.8)	15.3 (4.5)
Analytical Sec 6	16.5 (4.5)	16.0 (4.9)	14.7 (4.4)

Table 6.2 presents total score and section score intercorrelations for the CBT, S-T1GRE, and T1GRE samples. For ease in making comparisons among the samples, the table is structured as follows. The columns represent, for each sample, section scores and total scores by measure. For example, "V-5" refers to the score on verbal section 5; "A-T" refers to the total score on analytical. The rows represent the section and total scores by measure for each sample. Thus, "CBT-V Tot" refers to the verbal total score for the CBT sample, and "S-T1-Q Sec4" refers quantitative section 4 score for the S-T1GRE sample. Correlations are presented first for the CBT sample, followed by the S-T1GRE and T1GRE samples. These data indicate that the relationships among the section scores and the total scores are similar across the three samples. However, the correlations between verbal and analytical (both total and sections) are somewhat higher in the field test than in the October administration.

Table 6.2
Total Score and Section Score Intercorrelations
(decimals omitted)

Score	V-2	V-5	Q-T	Q-1	Q-4	A-T	A-3	A-6
CBT- V Tot	96	96	59	56	57	68	64	63
CBT- V Sec2		85	57	55	55	65	62	60
CBT- V Sec5			57	54	55	65	61	61
CBT- Q Tot				96	96	75	70	56
CBT- Q Sec1					85	73	68	68
CBT- Q Sec4						72	67	68
CBT- A Tot							94	94
CBT- A Sec3								77
S-T1 V Tot	97	97	53	52	51	68	66	64
S-T1- V Sec2		88	48	47	45	63	61	60
S-T1- V Sec5			56	54	54	69	68	65
S-T1- Q Tot				96	97	72	70	68
S-T1- Q Sec1					87	71	70	66
S-T1- Q Sec4						68	65	65
S-T1- A Tot							96	96
S-T1- A Sec3								83
T1- V Tot	96	95	58	55	56	63	60	58
T1- V Sec2		82	55	53	53	59	57	54
T1- V Sec5*			55	53	54	61	58	56
T1- Q Tot				96	96	74	69	69
T1- Q Sec1					86	71	67	66
T1- Q Sec4						71	67	66
T1- A Tot							94	93
T1- A Sec3								75

* Field test verbal section 5 was the seventh section in T1GRE.

6.3 Reliability and Standard Error of Measurement

Table 6.3 presents reliability coefficients and standard errors of measurement for each measure of the CBT, S-T1GRE, and T1GRE. The reliability of each score is based on the Kuder-Richardson formula 20 reliability coefficient for the measure. Table 6.3 indicates that the CBT produced reliability coefficients and standard errors of measurement comparable to those for the P&P versions.

Table 6.3
Reliabilities and Standard Errors of Measurement

		Reliability	Standard Error of Measurement
Verbal	CBT	.92	3.4
	S-T1	.92	3.5
	T1	.91	3.5
Quant	CBT	.92	3.0
	S-T1	.93	2.9
	T1	.92	3.1
Analytic	CBT	.89	2.8
	S-T1	.91	2.8
	T1	.88	2.9

6.4 Speededness

Indices of test speededness can be used to compare the CBT to S-T1GRE and T1GRE administrations. Section 8 describes additional CBT information regarding the number of items actually seen and answered. Using the ETS P&P definitions of items omitted and not reached for all three groups, Table 6.4 presents speededness information for each separately timed section. The four indices are:

- Percent Completing Section-- the percentage of examinees answering the last item of the section
- Percent Completing 75%-- the percentage of examinees answering at least 75% of the items
- Items Reached by 80%-- the number of items reached by 80% of the examinees
- NR Var. Divided by Score Var.-- Not-reached variance (i.e., the variance across examinees of the number of not-reached items) divided by the total score variance. If this ratio is .15 or less, the section is usually considered unspeeded for the group.

Table 6.4
Speededness Information

		Percent Answering Last Question	Percent Completing 75%	Items Reached by 80%	NR Var. Divided by Score Var.
Verbal Sec. 2	CBT	95	99	38	0.05
	S-T1	99	100	38	0.00
	T1	97	99+	38	0.03
Verbal Sec. 5	CBT	98	99+	38	0.02
	S-T1	97	100	38	0.00
	T1	98	99+	38	0.09
Quant. Sec. 1	CBT	86	99	30	0.06
	S-T1	99	100	30	0.00
	T1	94	99+	30	0.02
Quant. Sec. 4	CBT	88	99	30	0.05
	S-T1	97	100	30	0.02
	T1	96	99+	30	0.06
Analytic Sec. 3	CBT	86	99	25	0.07
	S-T1	99	100	25	0.01
	T1	95	99	25	0.07
Analytic Sec. 6	CBT	86	99	25	0.05
	S-T1	94	99	25	0.03
	T1	96	99	25	0.10

The first column of Table 6.4 suggests that the quantitative and analytical measures of the CBT may be slightly more speeded than either of the P&P versions of these measures. The percentages of CBT examinees completing the two verbal sections were about as high as the percentages for the other two samples. The percentages of examinees completing the quantitative and analytical sections of the CBT were smaller (by about 10%) than the corresponding percentages for the S-T1GRE and T1GRE groups. The speededness ratios for the CBT group were slightly larger than for the S-T1GRE group, but in only two of the sections were the CBT group speededness ratios larger than the ratios for the T1GRE group; however, the CBT speededness indices were still much smaller than the normal criterion indicating a speeded test. In addition, the speededness ratios for the CBT appear to drop slightly from the first to second section of each measure.

Note that the CBT and P&P versions were rights-scored tests; therefore, a common test-taking strategy is to answer every item because there is no penalty for wrong answers. However, it may be easier to rapidly fill in response ovals on a P&P answer sheet as the end of the test approaches than to click through uncompleted questions on the CBT. Thus, these indices of test speededness must be interpreted with caution, as they may reflect the test-taking strategy as much as test speededness.

Item response information in terms of distributions of number of rights, wrongs, omits, and not reached is presented for each test and measure in Table 6.5. Note that in this table the CBT omits and not reached are defined the same way as they are defined in the P&P versions. The mean numbers of not reached items were greater for the CBT than for the other two groups. This may suggest that the CBT may be slightly more speeded than the P&P version. On the other hand, it also may indicate that it is not as easy to click through and answer all items on the CBT as it is to fill in response ovals on the P&P answer sheet, and that examinees followed different test-taking strategies on the CBT and P&P versions.

Table 6.5
Distributions of Number Right, Wrong, Omit, Not Reached

		Number Right		Number Wrong		Number Omitted		Number Not Reached	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Verbal	CBT	49.9	12.0	25.4	11.7	0.3	1.2	0.4	2.0
	S-T1	49.8	12.2	25.9	12.1	0.2	0.7	0.0+	0.2
	T1	46.5	11.7	28.9	11.5	0.3	1.3	0.3	2.2
Quant	CBT	41.7	10.6	17.3	10.3	0.3	1.2	0.7	2.1
	S-T1	43.5	10.6	16.2	10.6	0.2	0.6	0.1	0.8
	T1	38.7	11.1	20.7	10.9	0.3	1.1	0.3	1.7
Analytic	CBT	33.8	8.4	15.0	8.4	0.5	1.6	0.7	1.8
	S-T1	32.7	9.3	16.7	9.4	0.5	2.2	0.2	1.2
	T1	30.0	8.4	19.2	8.3	0.4	1.6	0.4	2.3

6.5 Retest Effects Analyses

Table 6.6 shows differences in operational and field test number-right scores for each measure for examinees in the CBT and S-T1GRE groups. The upper half of the table shows percentages of examinees who increased their scores by specified interval amounts, and the lower half shows percentages of examinees who decreased their scores by specified interval amounts. For example, 52% of S-T1GRE examinees had higher verbal field test scores, as compared with operational scores, by as many as eight score points; and 30% of CBT examinees had lower analytical CBT scores, as compared with operational scores, by as many as six score points.

Table 6.6 shows that for the verbal and analytical measures, similar percentages of examinees in the CBT and S-T1GRE groups increased or decreased their scores by the specified amounts. For the quantitative measure, a somewhat greater percentage of the S-T1GRE group increased their scores than the CBT group; a correspondingly greater percentage of CBT examinees had lower scores at the field test. These data show that the CBT may have elicited slightly lower scores on the quantitative measure than did S-T1GRE. The mean (unadjusted) gains in number-right scores from operational to field test administrations (on the T1GRE number-right score metric) were as follows: for the S-T1GRE group, 0.93 (verbal), 1.29 (quantitative), and 1.41 (analytical); for the CBT group, 0.29 (verbal), 0.10 (quantitative), and 1.84 (analytical).

Table 6.6
Cumulative Distributions of Differences in Operational and Field Test Scores (in percents)

Score Difference*	Verbal		Quantitative		Analytical	
	CBT	S-T1GRE	CBT	S-T1GRE	CBT	S-T1GRE
0 - Total	53	58	52	66	64	63
0 - 10	50	54	50	61	59	60
0 - 8	47	52	47	56	55	56
0 - 6	40	43	42	51	45	48
0 - 4	29	32	33	40	31	37
0 - 2	15	20	18	22	17	17
0 - -2	15	15	18	11	16	13
0 - -4	26	25	32	22	26	21
0 - -6	36	33	39	28	30	28
0 - -8	41	38	45	33	34	31
0 - -10	44	41	47	34	35	35
0 - -Total	47	42	48	34	36	37

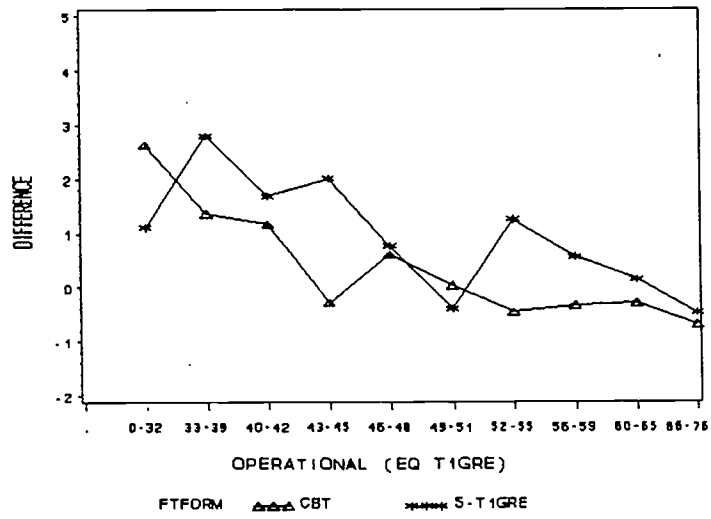
* Score difference is defined as field test number-right score (either CBT or S-T1GRE) minus equated T1GRE number-right score from the October 1991 administration.

Figure 6.4 shows the differences in operational and field test number right scores (in the T1GRE metric) by operational score for the CBT and S-T1GRE groups. The operational score has been divided into 10 levels where approximately one tenth of examinees fall into each category. The points on the plots are the average gain (loss) by examinees in the specified score-level category. The figure shows that for both groups and all three measures, score differences tended to decrease as operational scores increased. The plots for the verbal measure for the two groups overlapped in several places along the operational score scale. For the quantitative measure, the S-T1GRE group produced consistently higher score differences than the CBT group throughout the score range. For the analytical measure, the CBT group tended to have higher score differences than the S-T1GRE group throughout the score range; however, these differences between the groups were smaller than for the quantitative measure.

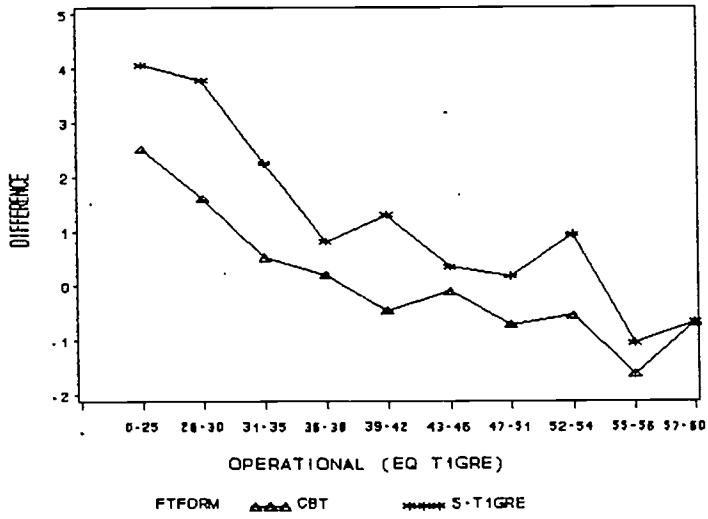
As both mode and retest effects could account for differences in operational and field test scores, another analysis was conducted to help isolate and evaluate these potential effects. Two primary hypotheses were tested: (a) mean scores for the groups did not change from operational test to field test and (b) any changes from operational to field test are the same for each group.

Figure 6.4
Field Test Minus Operational Number Right Score Differences by Operational Score Level

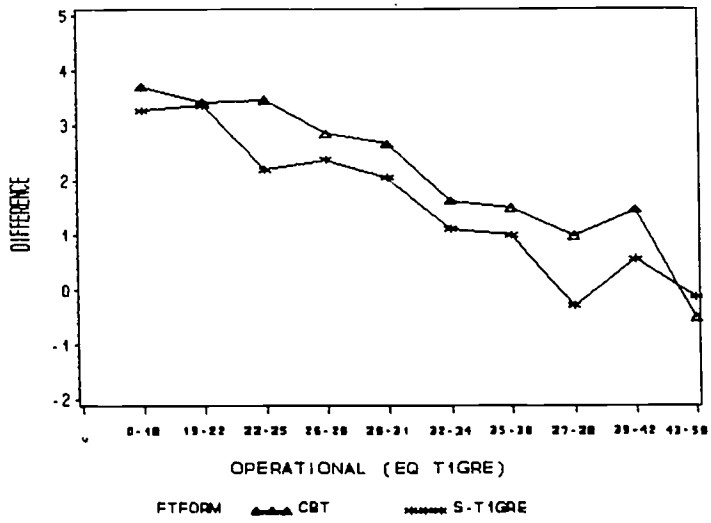
VERBAL



QUANTITATIVE



ANALYTICAL



To test the two hypotheses, an analysis of covariance (ANCOVA) was performed with difference score (field test score minus operational score put on the T1GRE metric) as the dependent variable, group (CBT or S-T1GRE) as the independent variable, and operational scores (put on the T1GRE metric) as the covariate.[5] Each measure was analyzed separately.

Table 6.7 shows the results of the ANCOVA analyses. The adjusted mean delta shows the mean change in scores adjusted for operational scores. It can be seen for the verbal measure that both groups gained on average less than one score point, and that the difference in gain between the groups was not statistically significant. For the analytical measure, both groups gained more than one score point, and these gains were statistically significant; however, the difference between the gains for the two groups was not significant.

For the quantitative measure, the S-T1GRE group gained more than one score point on average, and the CBT group gained about one tenth of a score point. A significant difference was found between the gains for the two groups.

Table 6.7
Results of Analysis of Covariance

Measure	Group	N	Score Means T2[6]	Score Means FT[7]	Adj. Mean Δ	Std. Error	95% C.I. Limits Lower	95% C.I. Limits Upper
V								
	S-T1	183	49.00	49.87	0.84	0.38	0.08	1.59
	CBT	1014	49.63	49.94	0.31	0.16	-0.01	0.63
Q[8]								
	S-T1	183	42.07	43.32	1.29[9]	0.33	0.64	1.94
	CBT	1014	41.61	41.73	0.11	0.14	-0.16	0.38
A								
	S-T1	183	31.13	32.50	1.27[9]	0.35	0.58	1.96
	CBT	1014	31.99	33.83	1.86[9]	0.15	1.57	2.15

[5] Form T2GRE number right scores were put on the T1GRE metric by equating T2GRE to T1GRE. In addition, as a test of the ANCOVA assumption of homogeneity of regression slopes within each group, an ANCOVA model was run that included a group by score interaction term. This analysis showed that the slopes were homogeneous. Finally, t-tests for each measure indicated that the two groups did not differ significantly in mean operational scores.

[6] T2 number right converted to T1 metric.

[7] FT = S-T1 for paper-and-pencil field test, CBT for CBT field test.

[8] Significant group effect for $\alpha = 0.01$.

[9] Significantly different from 0.0 at $\alpha = 0.01$.

One hypothesis regarding this potential quantitative test-level mode effect was that it occurred because the first quantitative section was also the first section of the CBT. A number of factors may have affected scores in the first section of the CBT. For example, immediately before beginning the first section of the CBT, examinees would have just read several introductory instruction screens and proceeded through the tutorials. Also, some examinees may have needed to go through the first section in order to become comfortable with the testing tools. If this were the case, scores on the first quantitative section of the CBT would be slightly depressed, as compared to the second quantitative section and to other sections of the test, when examinees would be more comfortable taking the CBT.

Table 6.8 shows the differences in mean section scores for each measure. (Mean section scores appear in Table 6.1.) In comparing mean section score differences between the CBT and S-T1GRE samples, respectively, the quantitative differences were 0.1 and 0.5, the verbal differences were 0.7 and 0.6, and the analytical differences were 0.8 and 0.6. (The first section of the quantitative and analytical measures elicited higher scores than the second section; the reverse was true for the verbal measure.) The biggest discrepancy between the two groups was the difference in quantitative section means. Note also that in the T1GRE sample the difference in quantitative section means was 0.3, which also was still larger than the 0.1 found for the CBT group. These data appear to provide some support for the hypothesis that scores in the first quantitative section of the CBT may have been slightly depressed because it was the first section of the test.

In evaluating these results, two points must be kept in mind. First, this slight effect did not noticeably affect individual items in the first section. Second, the effect is relatively small; that is, it appears to account for less than one half of a raw score point.

Table 6.8
Section Mean Score Differences

Mean Differences in Section Scores*			
Sample	Verbal	Quantitative	Analytical
CBT	-0.7	0.1	0.8
S-T1GRE	-0.6	0.5	0.6
Oct. '91 T1GRE	-0.4	0.3	0.6

* Reported as the first section mean minus the second section mean.

6.6 Conclusions

The data presented in this section suggest that the CBT is similar to the P&P versions in terms of score distributions, section score intercorrelations, test reliability, and standard error of measurement. There is some evidence of speededness on the CBT, where proportionally fewer CBT examinees finished the test than examinees in both P&P groups; however, the amount of any apparent speededness of the CBT is small. Also, these conclusions are based on ETS P&P definitions of omit and not reached. Test-taking behavior on computer may be different, and it is not clear what effect this would have on these indices.

No substantive test-level mode effects were found for the verbal and analytical measures. A slight test-level mode effect may be present in the quantitative measure that may, at least in part, be attributable to the fact that the first quantitative section was the first section of the CBT.

7.0 CBT Equating

The purpose of equating the CBT was to obtain raw-to-reported score conversions for the CBT. If no test-level mode effects were detected, the existing P&P score conversions would be applied to the CBT, and it would not be necessary to equate the form. However, differences were detected in the score distributions of the P&P and CBT forms that may have been due to factors other than only sampling error and retest effects. Therefore, it was necessary to equate the CBT to the existing P&P scale. The obtained conversions were then compared to the existing P&P conversions to determine if the CBT-based conversions were sufficiently different from the P&P conversions to warrant separate conversions.

7.1 IRT Equating

IRT true score equating (Lord, 1980) was performed to equate the CBT to the October 1991 administration of Form T1GRE. In this method, a raw score on the new form is assigned the reported score that corresponds to the same ability level on the old form (as measured by the θ-parameter in the IRT model). This method requires that both forms have IRT item parameter estimates on a common IRT scale (see Section 4 for a description of the IRT scaling process). Separate equatings were performed for each measure.

As mentioned in Section 4, the item characteristic curve method (Stocking & Lord, 1983) of common item scaling was used to scale each calibration to a base GRE scale. In performing the common item scaling, one decision regarded which items were to be treated as common items. Typically, when one or more items are replaced by new ones on a P&P test, the "new" form is equated back to its original version. This is done by treating the items that were not replaced as a common item block. In a analogous fashion, if any items were found to function differently on the CBT than in P&P, they could be not included in the common item block. Based on the conclusions of the item-level mode effects analyses, no items were excluded from the common item block in the primary equating analyses.

The resulting IRT true-score equatings for the CBT were compared to the original T1GRE conversions in two ways. First, the raw-to-reported score conversions were plotted on the same axes, allowing a direct visual inspection of the differences between them. Second, rounded reported score differences were computed and evaluated. These analyses were performed separately by measure.

7.2 Results

Figure 7.1 shows an overlay plot of the original raw-to-reported score conversion for Form T1GRE and the new raw-to-reported score conversion for the CBT for the verbal measure. Figures 7.2 and 7.3 show similar plots for the quantitative and analytical measures. As can be seen, the conversions for T1GRE and the CBT are virtually identical for all three measures.

Within the 200-800 score range, the rounded reported scores from the two conversions never differed by more than 10 reported score points (the smallest increment in the reported score metric) for any of the three measures. These differences were similar to those obtained when the T1GRE conversions were compared with the conversions obtained from equating the T1GRE criterion calibration sample (see Section 4.3) to the full T1GRE sample.[10] These results support the use of the T1GRE conversions for both the P&P and CBT versions of the form.

Appendix E shows difference plots of unrounded score conversions for T1GRE and the unrounded CBT IRT conversions for each measure at each number right score. To parallel GRE score reporting, converted scores less than 200 were set at 200 and converted scores greater than 800 were set at 800. These plots show that the differences were virtually always less than 10 scale score points, and almost always less than 5 scale score points. These differences are within the rounding done for reporting scores. The biggest differences came at the low end of the score scale, where the T1GRE conversions were higher than the CBT conversions (i.e., a negative score difference). However, only about 1% of CBT examinees scored at these levels.

In addition, additional IRT equatings were performed for each measure taking out the items flagged for b-parameter discrepancies from the common item block of the scaling process (see Section 5.1). The conversions from these equatings were almost identical to the IRT equatings that included all items in the scaling process, again suggesting that the items flagged were not of substantive interest.

7.3 Conclusions

IRT equatings of the CBT resulted in conversions that are virtually identical to the Form T1GRE conversions for each measure. As a result, it was recommended that the CBT have the same conversions as Form T1GRE. No adjustment was made for the slight test-level mode effect that may have occurred for the quantitative measure. Only test level mode effects that would occur throughout the duration of the administration of the CBT would require adjustments to the equating. For example, if the effect were due to familiarity with using a computer, and when the CBT becomes operational most examinees get more computer experience before taking the CBT, then adjusting the equating based on the field test results would be inappropriate. Also, due to the design of the field test, it was not possible to precisely identify whether this was a test-level mode effect and, if so, what its magnitude was.

Finally, as mentioned earlier, because of the CBT IRT scaling process, the similarity of equating conversion lines does not necessarily indicate the absence of main effects due to mode. However, other evidence (e.g., raw score distributions, ANCOVA results) points to that conclusion, particularly for the verbal and analytical measures. On the other hand, note that future CBT forms should not be assumed to have the same conversions as their P&P versions, even though the present CBT conversions closely matched the T1GRE conversions.

---

[10] To provide a baseline for evaluating the differences between the CBT and T1GRE conversions, the T1GRE criterion calibration sample was equated to the full T1GRE sample. The differences found between the criterion sample conversions and the original T1GRE conversions were due only to sampling errors. Thus, the differences in the CBT and T1GRE conversions were of magnitudes similar to those due to sampling errors.

Figure 7.1
IRT Equating Converted Score Comparison- Verbal

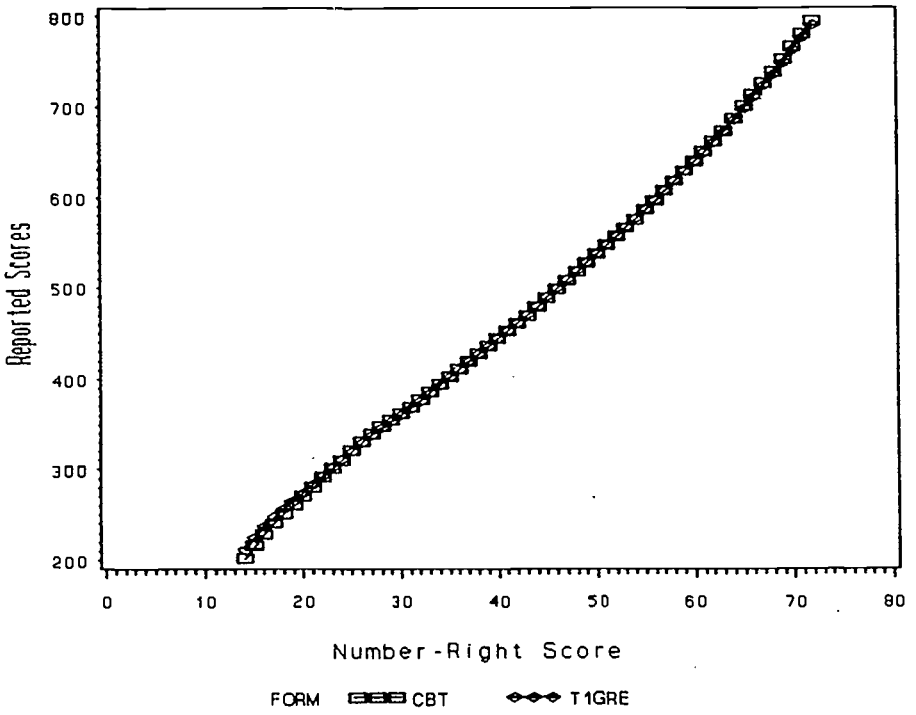


Figure 7.2
IRT Equating Converted Score Comparison- Quantitative

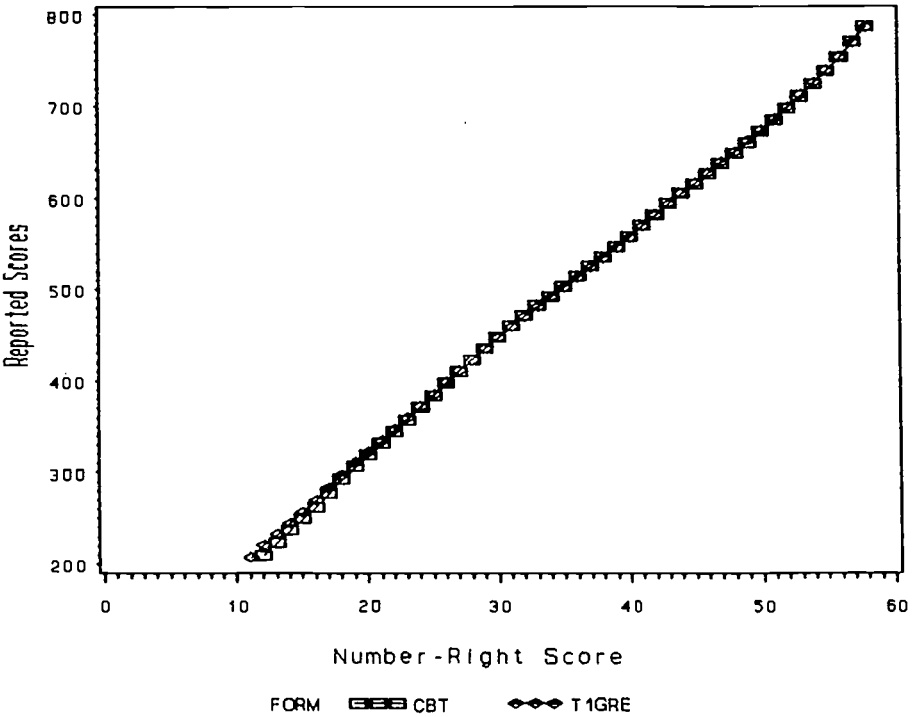
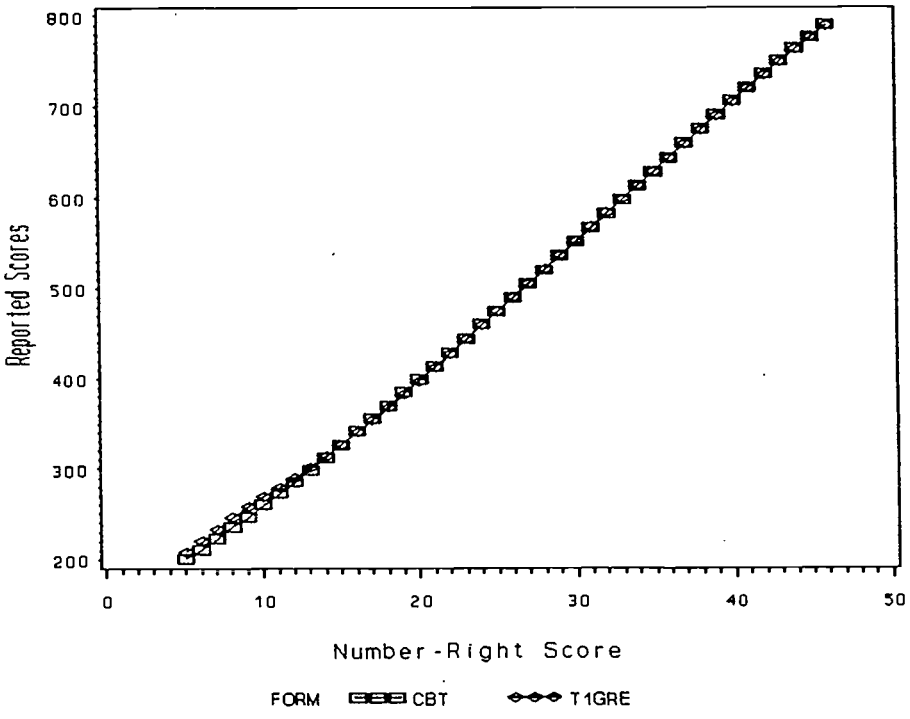


Figure 7.3
IRT Equating Converted Score Comparison- Analytical



8.0 CBT Timing

A test administration involves an interaction between an examinee and a set of test items. For a paper-and-pencil administration under standardized conditions, the examinee-items interaction can be described by the number of items (a) answered correctly, (b) answered incorrectly, (c) omitted, and (d) not reached. An omitted item is an unanswered item followed by one or more answered items. A not-reached item is an unanswered item followed only by unanswered items. The definitions of omitted and not reached assume that an examinee progresses through a test in a linear, sequential manner.

By administering a test via computer, more information on the examinee-items interaction was available. The information available for paper-and-pencil tests also was available; however, the definitions for omitted and not reached were refined to eliminate the linear progression assumption. On the CBT, an omitted item was defined as an item that appeared on the screen but was not answered by the examinee. A not-reached item was defined as an item that was not presented to the examinee (and thus not answered). In addition, the time spent on an item and the number of times an examinee visited an item were recorded and examined.

8.1 CBT Items Omitted and Not Reached

Timing data collected during CBT administrations allowed for definitions of omitted and not reached that did not assume examinees proceeded through the test sequentially. (A comparison of the numbers of omitted and not reached items for the P&P and CBT tests using paper-and-pencil definitions is presented in Section 6.4.) The revised definitions for omitted and not reached items, though an improvement over the paper-and-pencil definitions, are not without fault. An omitted item needs only to appear on the computer screen and not be answered. The definition does not account for items that appear on the screen for a short amount of time that is not long enough for the item to be processed by the examinee. Such an item would be classified as omitted but with a stricter definition taking time on screen into account would be considered a not reached item. Therefore, the definitions for omitted and not reached may overestimate the number of omitted items and underestimate the number of not-reached items.

Table 8.1 summarizes the number of omitted and not-reached items for each measure of the CBT. The mean number of not-reached items is less than or equal to half an item for each measure. Taking into account the number of items comprising each measure, the analytical measure was most speeded and the verbal measure least speeded. The mean numbers of omitted and not-reached items were similar. Given this information, the possible speededness problem discussed earlier is less of an issue than it appeared to be when a paper-and-pencil definition of not reached was used.

Table 8.1
Mean Number of Omitted and Not Reached Items

	Verbal	Quantitative	Analytical
# Omitted (seen but not answered)	0.3	0.5	0.7
# Not Reached (not seen)	0.3	0.5	0.5
# Items in Measure	76	60	50

8.2 Item Visits

The CBT delivery system allowed examinees to proceed nonsequentially through the test. Within a section, examinees were permitted to determine the order of item presentation and the number of visits to an item. The delivery system was designed to allow examinees a level of freedom to move about the test similar to that afforded examinees taking a paper-and-pencil test. In addition to recording item responses and item times, the CBT delivery system recorded the number of times an examinee visited each item. The number of items visited two or more times is summarized in Table 8.2.

Table 8.2
Mean (and Standard Deviation) of Item Visits

	Verbal	Quantitative	Analytical
# Items Visited Two or More Times	21.6 (15.2)	11.1 (10.3)	10.7 (9.00)
# Items in Measure	76	60	50

8.3 Help Visits

The CBT delivery system also allowed examinees to review material presented in the tutorials by invoking the Help feature. For the most part, examinees did not use the Help feature. On average, 0.2% of the examinees visited Help during any particular item. Approximately 60% of the examinees did not use Help at all, and another 34% used Help only once or twice. These data would suggest that the delivery system was accessible to the examinees and that the four tutorials presented at the beginning of the testing session were adequate in familiarizing examinees with the features and functions of the system.

8.4 Item Timing

The computer delivery of the items allowed for the collection of timing information that is not available for a paper-and-pencil test. The six operational sections of the CBT were separately timed with a time limit of 32 minutes for each section. For the most part, examinees used the full time allotted to complete a section. Therefore, if item times were averaged across measure with each measure being comprised of two sections, the averages would be expected to be approximately 47 seconds for verbal, 64 seconds for quantitative, and 77 seconds for analytical. A summary of the obtained item times is presented in Table 8.3. The actual mean times per item are similar to the values obtained by dividing the allotted time by the number of items in a measure.

Table 8.3
Mean (and Standard Deviation) of Item Times by Measure

	Verbal	Quantitative	Analytical
Item Times	44.8 (35.1)	57.3 (24.6)	72.5 (33.5)
# Items in Measure	76	60	50

A more informative approach to examining item timing would be to group items based on some defined dimension, such as difficulty or item type, and then examine the average amount of time per item with these groups. Examinees could also be grouped based on ability in order to examine the effect of ability on item timing.

For the present analyses, items were grouped on two dimensions--item difficulty and item type. Item difficulty was defined by using the equated deltas (see Section 5.2 for a description of deltas and delta equating) and items were classified as easy (deltas less than 10.9), middle (deltas between 11.0 and 14.0), and difficult (deltas greater than 14.0). The distribution of items, by measure, for the item difficulty levels is presented in Table 8.4.

Table 8.4
Distributions of Item Difficulty by Measure

	Item Difficulty		
Measure	Easy	Middle	Difficult
Verbal	25	38	13
Quantitative	26	24	10
Analytical	13	19	18

Item type was defined using broad content area categories. The item type categories and the number of items in each are presented in Table 8.5. Examinees were also grouped for the item timing analyses. Examinees were grouped separately for each measure into quintiles. The number-right scores on the CBT were used as indicators of ability.

Table 8.5
Distributions of Item Type by Measure

Measure	Item Type	# of items
Verbal	Sentence Completion	14
	Analogies	18
	Reading Comprehension	22
	Antonyms	22
Quantitative	Quantitative Comparison	30
	Discrete	20
	Data Interpretation	10
Analytical	Analytical Reasoning	38
	Logical Reasoning	12

Each measure of the GRE General Test contains some item sets (a group of items referring to common stimulus material). Given the layout of the test and the structure of the delivery system, the initial time spent on the stimulus material usually was recorded as time on the first item. For verbal and analytical, the average times for first items in sets was greater than the times for other items in sets or for discrete items. Therefore, it was decided to remove the first items in sets from all three measures. This translated to removing four verbal items, two quantitative items, and nine analytical items.

The first set of timing analyses examined the effect of item difficulty on item timing when conditioned on examinee ability. The mean item times in seconds for the verbal measure are presented in Table 8.6. Overall, examinees spent more time on middle items (47.5 seconds) than on difficult (40.5 seconds) and easy (24.2 seconds) items. The average time across items did not vary over ability groupings. Within an item difficulty category, average time spent on items did vary with examinee ability. For easy items, average time decreased as ability increased; for difficult items, average time increased as ability increased. The average time for middle items was somewhat constant across ability groupings. This can be better seen in the Figure 8.1.

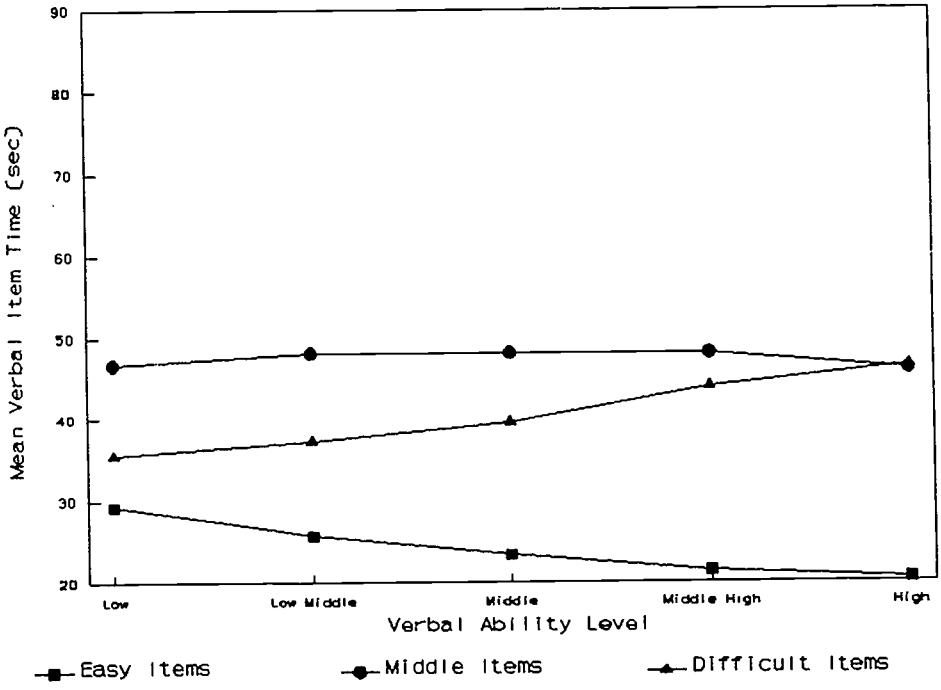
Table 8.6

Mean (and Standard Deviation) Item Time (in seconds)--Verbal By Examinee Ability and Item Difficulty (Excluding 4 Items That Appear First in Sets)

Ability	Item Difficulty			
CBT # Right	Easy	Middle	Difficult	Overall
Low	29.3 (6.1)	46.7 (8.0)	35.5 (8.7)	38.9 (6.0)
	25.7 (5.0)	48.2 (6.5)	37.3 (7.1)	38.7 (4.6)
	23.4 (4.5)	48.2 (6.3)	39.7 (7.6)	38.4 (4.4)
	21.5 (4.2)	48.1 (5.8)	44.0 (9.2)	38.5 (4.2)
High	20.6 (3.9)	46.1 (7.1)	46.5 (10.3)	37.7 (5.1)
Overall	24.2 (5.8)	47.5 (6.9)	40.5 (9.5)	

NOTE: Of the 4 items excluded, 1 item was easy and 3 were middle.

Figure 8.1
Mean Item Time (in seconds)--Verbal
By Examinee Ability and Item Difficulty
(Excluding 4 Items That Appear First in Sets)



Data for the quantitative and analytical measures showed similar patterns except that the average time per item increased as the difficulty of the items increased. Tables 8.7-8.8 and Figures 8.2-8.3 summarize the timing analyses by item difficulty for the quantitative and analytical measures.

The data presented for the timing analyses that focused on item difficulty should be interpreted with caution. Item difficulty is, to a large degree, confounded with item type. For example, 16 of the 38 middle verbal items are reading comprehension items[11]; and 10 of the 25 easy verbal items are antonyms. Differences among easy, middle, and difficult items may be partially due to the item type composition of the item difficulty groupings. However, the numbers of items available from the one form of the GRE General Test given as a CBT prohibits the examination of item difficulty and item type concurrently.

[11] Four of the 16 reading comprehension items were also first items in sets and were excluded from the analyses.

Table 8.7

Mean (and Standard Deviation) Item Time (in seconds)--Quantitative By Examinee Ability and Item Difficulty (Excluding 2 Items That Appear First in Sets)

Ability	Item Difficulty			
CBT # Right	Easy	Middle	Difficult	Overall
Low	51.5 (10.7)	64.0 (11.5)	53.2 (13.2)	57.0 (7.9)
	45.7 (7.8)	71.0 (8.1)	63.3 (14.7)	59.1 (4.5)
	42.1 (6.8)	72.7 (7.4)	69.2 (14.6)	59.4 (3.6)
	37.1 (5.6)	72.8 (7.0)	77.7 (13.8)	58.8 (4.1)
High	33.8 (5.5)	68.3 (9.0)	87.6 (17.2)	57.3 (5.7)
Overall	42.2 (9.9)	69.6 (9.4)	70.0 (19.1)	

NOTE: The 2 items excluded were both easy.

Figure 8.2
Mean Item Time (in seconds)--Quantitative By Examinee Ability and Item Difficulty (Excluding 2 Items That Appear First in Sets)

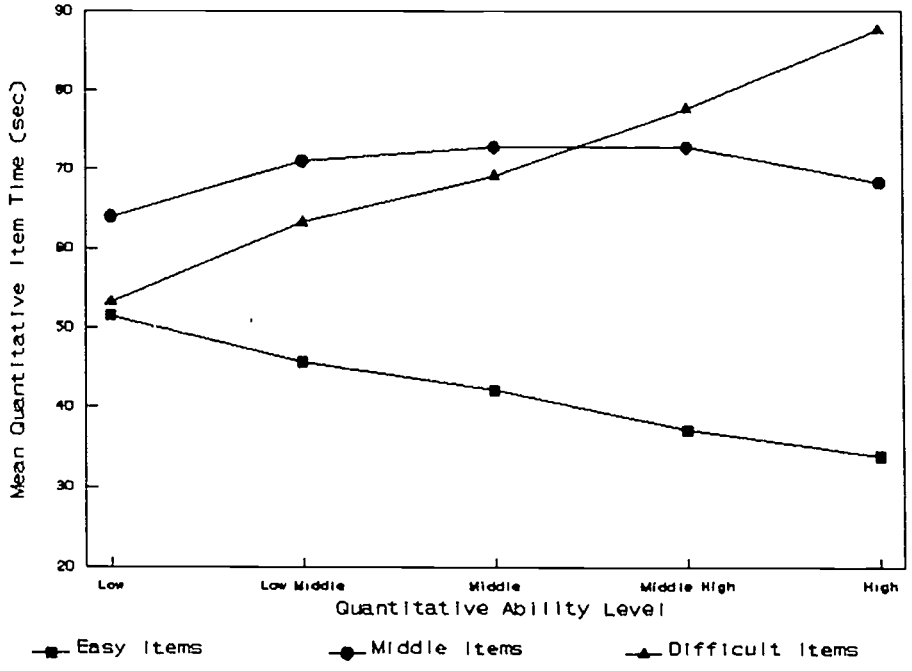


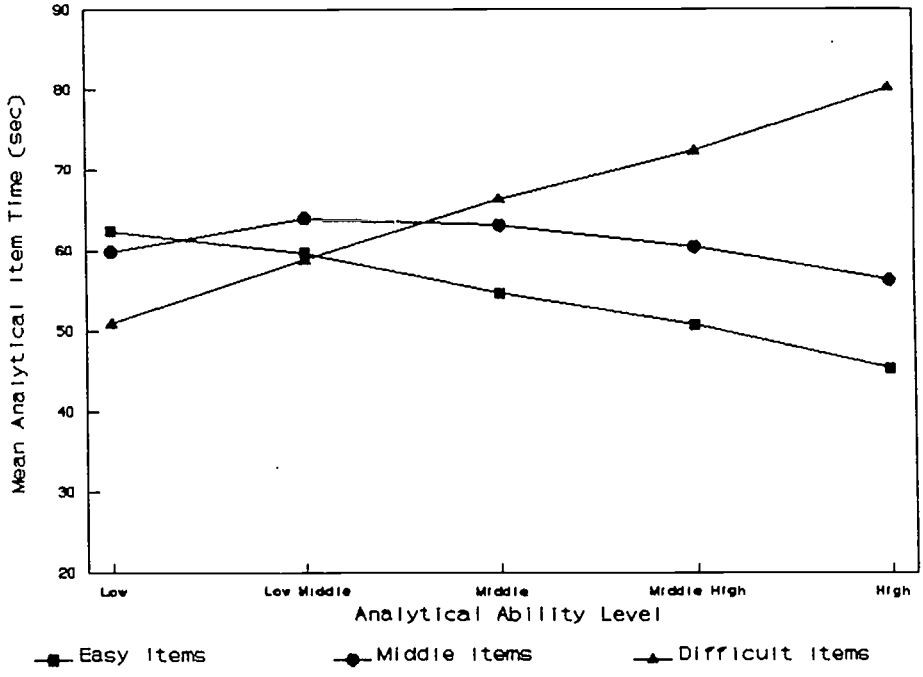
Table 8.8

Mean (and Standard Deviation) Item Time (in seconds)--Analytical Examinee Ability and Item Difficulty (Excluding 9 Items That Appear First in Sets)

Ability / CBT # Right	Item Difficulty: Easy	Middle	Difficult	Overall
Low	62.4 (19.8)	59.9 (13.9)	50.9 (13.0)	56.8 (11.1)
	59.7 (13.0)	64.0 (10.7)	58.8 (12.7)	60.9 (5.8)
	54.7 (10.2)	63.1 (11.0)	66.3 (13.2)	62.5 (5.6)
	50.7 (9.1)	60.4 (7.9)	72.3 (11.5)	63.1 (5.2)
High	45.3 (8.3)	56.3 (8.1)	80.1 (10.2)	63.7 (4.8)
Overall	54.5 (14.1)	60.7 (10.8)	65.8 (15.9)	

NOTE: Of the 9 items excluded, 4 were easy, 4 were middle, and 1 difficult.

Figure 8.3
Mean Item Time (in seconds)--Analytical By Examinee Ability and Item Difficulty (Excluding 9 Items That Appear First in Sets)



Similar analyses to those presented above for item difficulty were completed for item type. As can be seen in Table 8.9 for the verbal measure, large differences were found for average item times among item types. Reading comprehension items, on average, took more than two and a half times as long to answer as antonyms, except for sentence completion items, average times within an item type remained relatively stable over examinee ability groupings.

Table 8.9
Mean (and Standard Deviation) Item Time (in seconds)--Verbal
By Examinee Ability and Item Type
(Excluding 4 Items That Appear First in Sets)

Ability CBT # Right	Item Type			
	Sentence Completion	Analogies	Reading Comprehension	Antonyms
Low	49.5 (11.5)	33.9 (8.3)	55.5 (12.6)	21.5 (5.8)
	47.3 (9.0)	34.0 (8.2)	57.9 (11.5)	20.9 (5.3)
	44.0 (7.8)	34.0 (6.8)	59.6 (10.6)	20.9 (4.5)
	40.7 (8.3)	35.3 (7.2)	61.7 (10.3)	20.7 (5.0)
High	36.9 (8.4)	35.4 (7.6)	60.6 (11.2)	21.2 (5.3)
Overall	43.7 (10.2)	34.5 (7.7)	59.0 (11.5)	21.0 (5.2)

NOTE: All 4 excluded items were reading comprehension.

Tables 8.10 and 8.11 for the quantitative and analytical measures also show marked differences in average item times among the item types. For quantitative, discrete items take the longest amount of time, possibly due to the amount of unique stimulus material that must be incorporated in each item. As with the verbal measure, average times within a measure are fairly constant over ability groupings.

For the analytical measure, the overall difference between analytical reasoning and logical reasoning items is not substantially different (59.6 seconds compared to 66.5 seconds). However, unlike the case with verbal and quantitative, average times do differ across ability groupings. As ability increases, average times increase for analytical reasoning items and decrease for logical reasoning items. For low-ability examinees, the difference in the amount of time spent on the two item types is approximately 30 seconds.

Table 8.10
Mean (and Standard Deviation) Item Time (in seconds)--Quantitative
By Examinee Ability and Item Type
(Excluding 2 Items That Appear First in Sets)

Ability CBT # Right	Item Type		
	Quant. Comparison	Discrete	Data Interp.
Low	46.7 (11.6)	71.6 (16.0)	60.9 (16.5)
	46.8 (8.7)	75.6 (10.0)	65.3 (14.4)
	46.0 (7.4)	77.6 (8.6)	64.5 (11.7)
	45.2 (6.9)	77.0 (9.2)	64.6 (11.6)
High	44.3 (6.3)	74.3 (11.0)	63.8 (14.0)
Overall	45.8 (8.5)	75.2 (11.6)	63.8 (13.9)

NOTE: The 2 items excluded were both data interpretation.

Table 8.11
Mean (and Standard Deviation) Item Time (in seconds)--Analytical
By Examinee Ability and Item Type
(Excluding 9 Items That Appear First in Sets)

Ability CBT # Right	Item Type	
	Analytical Reasoning	Logical Reasoning
Low	49.1 (13.6)	78.7 (19.7)
	58.5 (8.7)	67.7 (14.8)
	61.9 (9.0)	63.8 (13.9)
	63.4 (8.4)	61.9 (11.7)
High	64.7 (7.2)	60.8 (11.0)
Overall	59.6 (11.0)	66.5 (15.8)

NOTE: The 9 items excluded were 8 analytical reasoning and 1 logical reasoning.

9.0 CBT Questionnaire Results

At the end of the CBT administration, examinees were asked to complete a questionnaire. The questionnaire covered a variety of topics, including prior computer experience, specific reactions to the CBT, and CBT or P&P comparisons and preferences. All CBT examinees in the sample completed the questionnaire. A copy of the questionnaire is in Appendix F.

The percentage of respondents selecting the different alternatives to each question appears next to the question number on the questionnaire in Appendix F.[12] For example, 45% of respondents indicated in question 1 that they regularly used a personal computer, and in question 33, 60% of respondents indicated that they would prefer taking a computer test over a paper-and-pencil test with the same questions. The percentages corresponding to open-ended questions refer to the percentages of examinees who provided any comments. For example, in question 2, 16% of examinees listed one or more programming languages. Percentages are computed relative to the total number of respondents to the questionnaire.

Table F.1 in Appendix F lists percentages of examinees from gender and ethnic subgroups, as well as the total group, who selected each alternative to each question. For example, 38% of Black respondents indicated in question 2 that they used a personal computer for graphics; in question 13, a total of 63% of female respondents found the time display helpful during the last five minutes of each section.

Several issues were of concern going into the field test, two of which are described here. One was that examinees who did not use computers regularly or who did not have access to computers would feel disadvantaged taking the CBT. The majority of field test participants reported using computers regularly (81%), and most reported they owned or had access to a computer where they lived (68%). For the 31% of examinees that reported they did not own or have access to a computer, 57% indicated they would prefer to take a CBT version rather than a P&P version; 32% would prefer P&P. Approximately 9% expressed no preference. These percentages are similar to those for examinees who owned or had access to computers where they lived. This information is presented in Table 9.1. Given this information, the concern that examinees with limited access to computers would feel disadvantaged is unsupported.

Table 9.1
Examinee Preference Conditioned on Computer Access*

Computer Access (Question 3)	Preference (Question 33)		
	CBT	P&P	No Preference
YES	61%	26%	11%
NO	57%	32%	9%

* Percentages indicate the percentages of examinees in a row preferring CBT, P&P, or neither.

[12] Responses provided by examinees to the open-ended questions are not listed in this report; however, the percentage of examinees providing a response to each open-ended question is provided.

Another concern was that the extensive use of the mouse to navigate through the test would be a disadvantage to those who did not use a mouse regularly. Results of the questionnaire showed that, on the whole, examinees felt they were able to navigate through the test using the mouse. In addition, for the 50% of the examinees who reported they never or rarely used a mouse before the field test, 58% said they would prefer a CBT version of a test over a P&P version; 30% said they would prefer a P&P version. Approximately 10% expressed no preference. The percentages are similar to those for examinees who reported they regularly or routinely used a mouse. This information is presented in Table 9.2. Examinees with little or no exposure to a mouse did not perceive any disadvantage with using it to navigate through the CBT delivery system.

Table 9.2
Examinee Preference Conditioned on Mouse Use

Mouse Use (Question 10)	Preference (Question 33)		
	CBT	P&P	No Preference
Regularly or Routinely	61%	25%	11%
Rarely or Never	58%	30%	10%

* Percentages indicate the percentages of examinees in a row preferring CBT, P&P, or neither.

As can be seen from responses throughout the questionnaire, the consensus was very favorable to the CBT. The vast majority of examinees reported having little or no trouble navigating through the test. Most examinees had either no preference or preferred the CBT over a P&P test. Most examinees were frustrated when they were not allowed to use the Review option during the seventh section. Also, the subgroup analyses of the questionnaire revealed the groups were similar in their responses to the questions.

10.0 Subgroup Analyses

The sample of examinees obtained was similar to the GRE national population in terms of ethnic and gender proportions. The subgroup sample sizes were sufficient to provide some meaningful descriptive statistics, although larger sample sizes would be required for more thorough analyses. CBT timing data are presented below for gender and ethnic subgroups. Subgroup means and standard deviations on the national test and on the field test forms are provided for the CBT and S-T1GRE groups. Also, DIF results by gender are presented.

10.1 Subgroup CBT Timing Information

Table 10.1 shows the mean times spent on each tutorial by gender and ethnic subgroup as well as for the total group. Somewhat less than 20 minutes was spent on average by each group on all tutorials. The differences between female and male examinees were small, and Black and Hispanic examinees spent slightly more time on the tutorials than did other subgroups.

Table 10.1
Mean Tutorial Times by Subgroup

Tutorial	Mean Times in Minutes						
	Female	Male	Asian	Black	Hispanic	White	Total
Mouse	3.3	3.1	2.8	3.7	3.8	3.1	3.2
Testing tools	7.7	7.4	6.9	8.8	7.9	7.4	7.5
Answering items	2.5	2.3	2.1	3.0	2.7	2.3	2.4
Scrolling	3.9	3.4	3.3	4.1	4.1	3.6	3.7
Total tutorial	17.3	16.2	15.1	19.8	18.5	16.5	16.8
N. Examinees	617	397	65	80	50	782	1,017

In addition, for each subgroup the average number of Help visits was less than one and the amount of time using the item Help function was less than 10 seconds on average.

Table 10.2 shows the mean number of items seen and answered on the CBT for each subgroup. An item was considered to have been seen if it appeared on the examinee's CBT screen, regardless of whether it or subsequent items were answered. Section 8 provides additional detail on the definitions of these variables. These data provide evidence about whether examinees were able to navigate through the test and use the test-taking strategy of answering all items. The means indicate that all groups saw and answered almost every item. Black examinees are a possible exception: for each measure, they saw and answered about one less item on average than other examinees. For the total group, the mean numbers of items answered are comparable to those found in P&P operational administrations.

Table 10.2
CBT Mean Number of Items Seen and Items Answered by Subgroup

	Means						
	Female	Male	Asian	Black	Hisp.	White	Total
Verbal Items Seen	75.7	75.5	75.9	74.4	75.2	75.8	75.7
Verbal Items Answered	75.4	75.2	75.8	73.7	74.9	75.5	75.4
Quant Items Seen	59.5	59.6	59.9	58.8	59.6	59.6	59.5
Quant Items Answered	58.8	59.3	59.7	57.8	59.0	59.1	59.0
Analyt Items Seen	49.4	49.6	49.7	49.0	49.4	49.5	49.5
Analyt Items Answered	48.7	48.9	49.2	48.0	49.0	48.8	48.8
N. Examinees	617	397	65	80	50	782	1,017

10.2 Subgroup Score Information

Table 10.3 lists the means and standard deviations of CBT and national scores for subgroups. These data are provided for purposes of describing the CBT sample. Table 10.4 lists the means and standard deviations of field test and national scores by gender for the S-T1GRE sample. The study was not designed to investigate subgroup score differences and subgroup sample sizes were small; thus, no generalizable inferences can be drawn from these data.

Table 10.3
Subgroup Score Means (and Standard Deviations) for the CBT Group

	Female	Male	Asian	Black	Hispanic	White	Total
CBT Verbal	49.0 (11.6)	51.4 (12.5)	52.0 (12.3)	38.8 (9.4)	45.5 (13.5)	51.2 (11.6)	49.9 (12.0)
Nat. Verbal*	48.5 (11.8)	51.4 (11.5)	52.8 (11.3)	37.7 (9.9)	46.1 (12.4)	50.9 (11.2)	49.6 (11.8)
CBT Quant	39.1 (10.3)	45.9 (9.8)	48.4 (8.0)	32.4 (11.2)	38.6 (9.8)	42.4 (10.2)	41.7 (10.6)
Nat. Quant*	39.4 (10.4)	45.1 (10.1)	48.9 (7.8)	32.2 (10.9)	37.7 (9.8)	42.3 (10.2)	41.6 (10.6)
CBT Analytic	33.0 (8.3)	35.1 (8.5)	36.7 (7.6)	25.1 (7.9)	29.1 (8.2)	34.8 (7.9)	33.8 (8.4)
Nat. Analyt*	31.4 (8.1)	32.9 (8.2)	33.6 (8.0)	23.1 (6.7)	27.8 (7.7)	33.1 (7.7)	32.0 (8.1)
N. Examinees	617	397	65	80	50	782	1,017

* National scores are on the T1GRE metric.

Table 10.4
Score Means (and Standard Deviations) by Gender for the S-T1GRE Group

	Female	Male	Total
S-T1GRE Verbal	49.5 (12.0)	50.4 (12.6)	49.9 (12.2)
Nat. Verbal*	47.8 (11.9)	50.4 (11.9)	49.0 (12.0)
S-T1GRE Quant	40.7 (10.0)	46.4 (10.8)	43.4 (10.7)
Nat. Quant*	39.5 (10.7)	45.0 (11.7)	42.1 (11.5)
S-T1GRE Analytic	32.3 (8.7)	32.8 (10.2)	32.5 (9.4)
Nat. Analyt*	30.9 (8.5)	31.3 (9.6)	31.1 (9.0)
N. Examinees	100	83	184

* National scores are on the T1GRE metric.

10.3 Gender Differential Item Functioning

Mantel-Haenszel (MH) gender analyses were conducted on the CBT data to examine DIF (sample sizes of the ethnic groups were too small for ethnic group analyses). Only one item was flagged as a C item, and it indicated DIF of a verbal item favoring female examinees.

10.4 Conclusions

The CBT timing information revealed no large differences between gender or ethnic subgroups. All subgroups spent somewhat less than 20 minutes on average on the tutorials. Most subgroups saw, for each measure on average, all but about one-half item and answered all but about one item. The average numbers of items seen and answered by Black examinees were slightly less than for the other subgroups; however, P&P comparisons of subgroups on the number of items seen are not possible. Descriptive statistics on national and CBT scores were provided. Finally, MH gender analyses yielded no items that exhibited DIF against female examinees and one verbal item that exhibited DIF in favor of female examinees.

11.0 Conclusions

Based on the results presented in this report, several major conclusions can be drawn. These conclusions focus on (1) examinees' interaction with and reaction to the CBT, (2) item- and test-level mode effects, and (3) placing the CBT on the existing GRE reporting scale.

11.1 Examinees' Interaction and Reaction to the CBT

Data from the field test indicate that examinees were able to navigate through the test with very little difficulty. Examinees reported little difficulty with using the mouse, Help, or Review. The majority of examinees used both Previous and Review to return to unanswered and/or marked questions, which indicates nonsequential progression through the CBT similar to P&P test-taking strategies.

The fact that few examinees used Help also may support the conclusion that the delivery system was accessible. The few examinees that who did use Help used the facility on only one or two items out of the 186 operational items of the CBT. Given the self-reported and Help visit data, it appears the test delivery system was comfortable for examinees.

Evidence from the field test also shows that the 32-minute-per-section time limits were sufficient. The average time per item for each measure was slightly less than the allotted time. However, fewer examinees answered the last question in the quantitative and analytical sections than answered the last question in the sections on the P&P test. Further insights into the extent of this speededness question may be gained with future examination of the item timing data.

Finally, the examinees' overall reaction to the CBT was favorable. When asked how they would rate the computer test-taking experience as compared with that of the P&P test, the majority rated the CBT better than the P&P test or about the same.

11.2 Item- and Test-Level Mode Effects

Item-level mode effects were assessed using IRT, classical statistics, Mantel-Haenszel (DIF) analyses, and logistic regression analyses. Differences in classical item difficulty and discrimination statistics were minor. Differences in IRT statistics were also minor. Mantel-Haenszel analyses flagged only one item for each measure (two of which favored the CBT group). Similarly, logistic regression yielded no substantive item-level mode effects.

Comparison of score gains on the CBT (over the October regular administration) with those of a P&P retest indicated that similar percentages of examinees increased or decreased their scores by similar amounts on the verbal and analytical measures. A small test-level mode effect was found for the quantitative measure. This effect was evidenced by a roughly one raw score point smaller increase in scores for examinees who took the CBT than for those who took the P&P field test (one raw score point would represent a scaled score difference of ten or fewer points). Given that the source of the test-level effect is unclear, it is advisable to continue to investigate this issue.

11.3 Placing the CBT on the GRE Reporting Scale

IRT true score equating was used to equate the CBT to the original paper-and-pencil test. Visual inspection of overlay plots of raw-to-reported score conversions for each of the three General Test measures indicates that the CBT and P&P conversion lines are virtually identical. Rounded reported scores from the two conversions never differ by more than 10 scaled score points (the smallest scaled score difference). The magnitude of the differences is similar to that expected by sampling error. The results support the use of the same conversions for the P&P and CBT versions of the test.

References

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity. Hillsdale, NJ: Erlbaum.

Kingston, N. M., & Dorans, N. J. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. Applied Psychological Measurement, 9, 281-288.

Lord. F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 4, 361-370.

Wingersky, M. S., Patrick, R. P., & Lord, F. M. (1988). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Appendix A
Times Spent on Tutorials

Table A.1 summarizes the results of the analyses of tutorial times. As can be seen from this table, the average total tutorial time was 16.8 minutes, and the range was from 5.9 minutes to 48.5 minutes.

Table A.1
Tutorial Times--Summary Statistics in Minutes

		Tutorial Type			
Statistics	Total	Mouse	Tools	Answering	Scrolling
Mean	16.8	3.2	7.5	2.4	3.7
S.D.	5.7	1.3	2.6	1.0	1.9
Minimun	5.9	0.8	3.0	0.8	0.6
Maximum	48.5	12.0	25.0	8.3	20.8

Appendix B
IRT Parameter Estimates for the CBT

Verbal

Section 2				Section 5			
Item	a	b	c	Item	a	b	c
1	0.492998	-6.513733	0.138400	1	0.819943	-3.559662	0.138400
2	0.835728	-0.524104	0.179613	2	0.346094	-1.819287	0.138400
3	0.686133	-0.784502	0.000000	3	0.819980	-1.644837	0.000000
4	1.013188	-0.985293	0.080890	4	0.517297	-0.777129	0.138400
5	0.819463	-0.402947	0.000000	5	1.338869	0.340812	0.390831
6	0.897294	0.921712	0.364469	6	0.739939	0.306442	0.082801
7	1.318631	0.378431	0.120110	7	0.713749	-0.018439	0.119926
8	0.732165	-3.894444	0.138400	8	0.530187	-3.076338	0.138400
9	0.558781	-3.851908	0.138400	9	0.403170	-2.284055	0.138400
10	0.368563	-2.753516	0.138400	10	0.770126	-2.043913	0.138400
11	0.593676	-0.424378	0.260787	11	0.457483	-1.361520	0.138400
12	0.834228	-0.297457	0.095342	12	1.069150	-0.698776	0.161953
13	0.880191	0.418058	0.084619	13	0.448407	1.108521	0.184542
14	0.564282	0.850368	0.091603	14	0.638089	1.217316	0.129184
15	0.703714	1.326216	0.266565	15	0.793233	0.741022	0.335321
16	0.824084	1.403960	0.100098	16	0.627967	0.476372	0.000000
17	0.560696	-0.661186	0.106428	17	1.116498	-0.236839	0.362196
18	0.931151	-1.304281	0.296986	18	0.628334	-1.572590	0.138400
19	0.634564	-0.503267	0.225714	19	0.836464	0.080330	0.087783
20	0.673267	-0.654621	0.251974	20	0.698093	-0.905469	0.115879
21	0.315612	1.045963	0.078155	21	0.582912	-0.611462	0.138400
22	0.509136	1.247513	0.130072	22	1.052315	0.653380	0.224513
23	0.405920	0.386142	0.021163	23	0.375124	0.730859	0.058046
24	0.697827	-1.224127	0.077313	24	0.438976	0.129760	0.030272
25	0.786640	0.056164	0.031520	25	0.575042	-0.966214	0.138400
26	0.934321	-0.193354	0.169404	26	1.014324	1.212779	0.205825
27	0.820579	-0.095558	0.105393	27	0.486117	0.588558	0.243780
28	0.931432	-2.516946	0.138400	28	0.368188	-6.245996	0.138400
29	0.558345	-3.373557	0.138400	29	1.224996	-1.683276	0.500000
30	1.451487	-0.712325	0.500000	30	0.327639	-4.935006	0.138400
31	1.089746	-0.984143	0.429934	31	0.684072	-2.963896	0.138400
32	0.762548	-0.937635	0.130667	32	0.924255	-1.107442	0.021012
33	1.128424	0.360900	0.159722	33	0.708900	0.409783	0.217088
34	1.046103	0.875543	0.224186	34	0.818544	0.674227	0.098727
35	0.681920	0.447450	0.026287	35	0.559289	0.355237	0.084766
36	1.451487	0.921513	0.033006	36	1.257319	1.350818	0.229949
37	1.451487	1.613139	0.093853	37	1.119370	1.650147	0.085864
38	0.778496	2.160360	0.001367	38	1.040755	1.736918	0.000000

Appendix B (continued)
IRT Parameter Estimates for the CBT

Quantitative

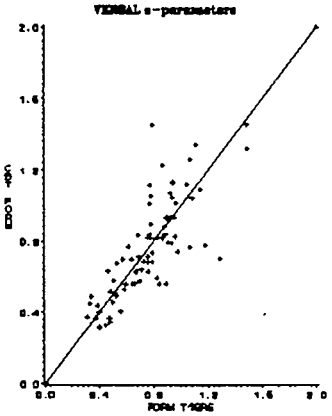
Section 1				Section 4			
Item	a	b	c	Item	a	b	c
1	0.567113	-2.842360	0.188598	1	0.491808	-2.856698	0.188598
2	0.526402	-2.009890	0.188598	2	0.416443	-2.842957	0.188598
3	0.809515	-1.552770	0.006859	3	0.969555	-1.575844	0.000000
4	0.990714	-0.223359	0.500000	4	0.863854	-0.884882	0.338609
5	0.925766	-0.937803	0.000000	5	1.167373	-1.058507	0.354478
6	0.881458	-1.236933	0.000000	6	1.415687	-1.014811	0.241712
7	0.857986	-1.175902	0.000000	7	1.188334	0.275545	0.500000
8	0.963131	-0.005980	0.314490	8	1.084549	-0.672427	0.118116
9	0.620024	-1.894977	0.188598	9	1.471969	-0.610972	0.213230
10	0.657054	-0.554903	0.075379	10	0.884582	-0.608975	0.043033
11	1.815852	1.057028	0.480709	11	1.201369	0.669335	0.335323
12	1.815852	-0.026409	0.292950	12	1.574074	0.453690	0.289292
13	1.496390	-0.018566	0.178471	13	1.242927	0.341004	0.124093
14	0.666004	1.421324	0.127670	14	0.923663	1.216927	0.074302
15	0.886404	0.974591	0.000000	15	1.252505	1.246198	0.139105
16	0.607495	-1.765325	0.119776	16	0.785038	-0.956055	0.206309
17	1.212263	-1.302910	0.000000	17	0.964584	-1.071389	0.231171
18	1.292901	-0.415792	0.137226	18	0.779474	-1.100677	0.000000
19	0.870373	-0.193043	0.073621	19	1.143773	-0.602190	0.077366
20	1.113927	0.345936	0.300338	20	0.809929	0.379701	0.229351
21	0.398851	-3.633798	0.119776	21	0.659396	-3.104693	0.119776
22	0.662114	-1.474798	0.119776	22	1.198719	-1.407638	0.000000
23	0.575048	-2.060522	0.119776	23	1.000507	-0.399830	0.217881
24	0.671826	-0.451364	0.000000	24	1.015346	-0.547570	0.000000
25	0.592207	1.200152	0.140229	25	1.363057	0.875162	0.099064
26	1.570288	-0.433323	0.063705	26	0.910089	0.356041	0.089287
27	1.815852	0.600538	0.205961	27	1.815852	0.198491	0.357494
28	1.496222	0.707690	0.088318	28	1.508181	0.733458	0.261799
29	1.393213	0.831192	0.123476	29	0.547776	0.603359	0.000000
30	1.815852	0.710563	0.187886	30	0.814198	1.049329	0.115001

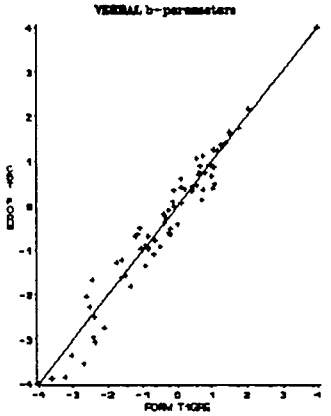
Appendix B (continued)
IRT Parameter Estimates for the CBT

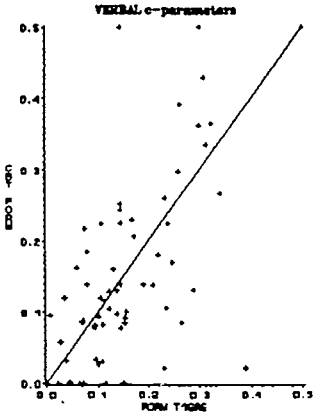
Analytical

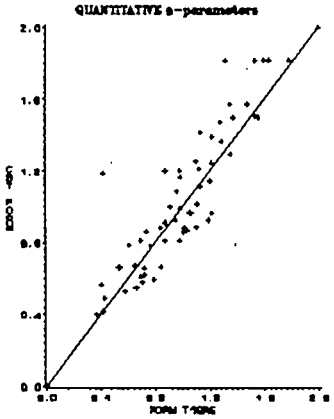
Section 3				Section 6			
Item	a	b	c	Item	a	b	c
1	0.615299	-3.578787	0.130505	1	0.781383	-1.810510	0.130505
2	0.569728	-1.008602	0.130505	2	0.568764	-0.938679	0.130505
3	0.538566	-0.233166	0.000000	3	1.171886	-0.733711	0.056080
4	0.567648	-1.664544	0.130505	4	0.965634	-0.856873	0.261846
5	0.648921	0.585476	0.075498	5	0.959734	-1.271963	0.021444
6	1.346742	-0.836960	0.083813	6	0.914259	-1.635044	0.130505
7	0.334010	-4.686453	0.130505	7	0.732666	-1.908843	0.130505
8	0.625029	-0.153573	0.436194	8	0.713802	-0.152665	0.209723
9	0.473872	-0.042093	0.315658	9	0.557265	-0.955581	0.130505
10	0.670661	-2.357620	0.130505	10	0.840347	-0.970824	0.000000
11	1.104050	-1.207949	0.307549	11	0.869724	0.662865	0.172200
12	0.718547	-0.833381	0.000000	12	0.778161	-1.315626	0.130505
13	0.855225	-0.871456	0.109234	13	0.725115	-1.244563	0.130505
14	0.773865	1.014771	0.111025	14	0.852332	-0.496133	0.169468
15	0.708203	-0.399813	0.198531	15	0.986604	0.937636	0.180248
16	1.179865	0.016350	0.175230	16	1.007210	0.105014	0.291425
17	1.194984	-0.215768	0.153436	17	0.990737	0.832243	0.108924
18	1.232376	0.141362	0.280692	18	0.919936	0.834267	0.176726
19	0.859090	-0.229797	0.253302	19	1.219779	1.233999	0.167024
20	1.614989	1.383684	0.149687	20	0.737978	1.121147	0.145561
21	0.929296	1.025769	0.223390	21	1.242906	1.400518	0.223302
22	1.535620	1.333980	0.089712	22	1.068202	1.632081	0.179232
23	0.546436	0.707215	0.190032	23	0.711242	1.137943	0.423784
24	0.717426	1.159048	0.320200	24	0.418787	1.661985	0.130505
25	0.690107	1.719311	0.269327	25	0.378396	2.731391	0.038216

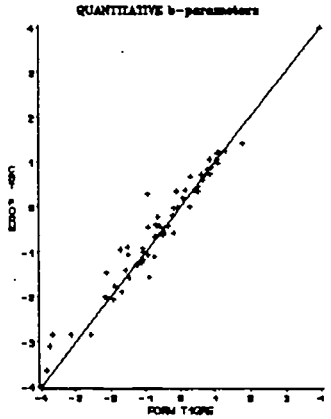
Appendix B (continued)
Plot of CBT and T1GRE IRT Parameter Estimates

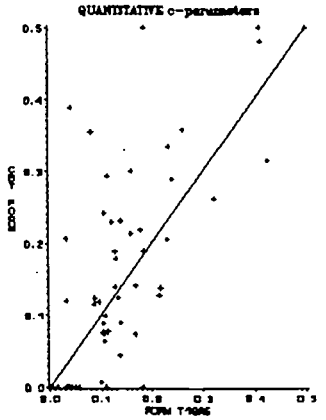


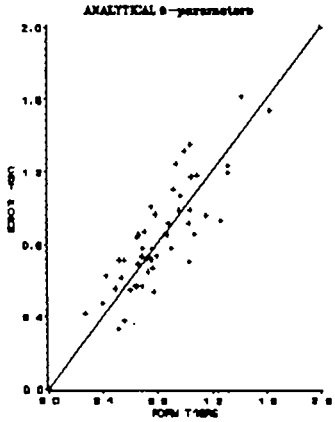


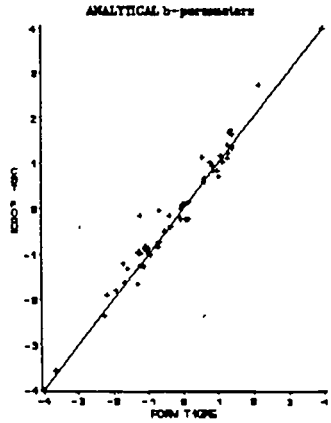


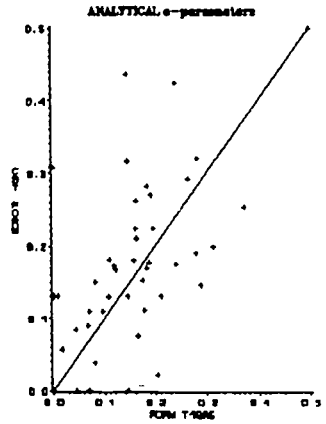












Appendix C
IRT Item Flagging Process: Technical Details and Results

Comparisons of item parameter estimates were performed as follows. First, for each parameter (a,b and c), the differences (d) between the values obtained for the CBT (or the criterion calibration) and T1GRE were computed. Then, the maximum negative difference (Max D-), the maximum positive difference (Max D+), and the mean difference (MD) were determined. In addition, a root mean squared difference (RMSD) between the two sets of estimates was computed for each parameter, as was a mean absolute difference (MAD). The RMSD statistic was computed as

$$RMSD_x = \left(\frac{1}{n}\sum_{i=1}^{n}(x_{1i}-x_{2i})^2\right)^{1/2}, \quad (2)$$

where x is either the a-parameter, the b-parameter, or the c-parameter, $RMSD_x$ is the RMSD statistic for parameter x, x_{1i} is the estimate of the item parameter from the T1GRE calibration, x_{2i} is the estimate from the CBT (or criterion) calibration, and n is the number of items. The MAD statistic is given by

$$MAD_x = \frac{1}{n}\sum_{i=1}^{n}|x_{1i}-x_{2i}|, \quad (3)$$

where MAD_x is the MAD statistic for parameter x, and the remaining terms are as previously defined.

Comparisons of the ICCs proceeded as follows. First, for each estimated ability (θ_j) in the T1GRE calibration, the difference in the probability of a correct response based on the CBT (or criterion) calibration item parameter estimates and the T1GRE item parameter estimates were computed as:

$$D_{ij} = P_{1i}(\hat{\theta}_j) - P_{2i}(\hat{\theta}_j), \quad (4)$$

where $P_{1i}(\theta_j)$ is the probability of a correct response to item i for an examinee of ability θ_j using the CBT (or criterion) calibration item parameter estimates and equation 1, $P_{2i}(\theta_j)$ is the probability of a correct response to item i for an examinee of ability θ_j using the T1GRE calibration item parameter estimates and equation 1, and D_{ij} is the difference in probabilities for item i for ability j.

Next, for each item the ICC mean difference (IMD) across all N_j estimated abilities was computed as:

$$IMD_i = \frac{1}{N_j}\sum_{j=1}^{N_j} D_{ij}. \quad (5)$$

Then the item with the largest negative ICC mean difference (Max IMD-) and the item with the maximum positive ICC mean difference (Max IMD+) were identified, and the mean over items of the ICC mean differences (MIMD) was computed. The MIMD statistic was computed as:

$$MIMD = \frac{1}{n}\sum_{i=1}^{n} \bar{D}_i \, . \tag{6}$$

Following this, the root mean (over items and examinees) squared difference (RMMSD) was computed as:

$$RMMSD = \left(\frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}(P_{1ij}-P_{2ij})^2\right)^{1/2} , \tag{7}$$

and the mean over items and examinees of the absolute differences (MMAD) in probabilities was computed as:

$$MMAD = \frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}|P_{1ij}-P_{2ij}| \, . \tag{8}$$

Appendix C (continued)

Table C.1

Calibration Differences by Comparison Type and Measure

		Verbal		Quantitative		Analytical	
		CBT	Crit.[13]	CBT	Crit.[11]	CBT	Crit.[11]
a							
	Max D-	-0.579	-0.245	-0.268	-0.379	-0.336	-0.364
	Max D+	0.657	0.442	0.769	0.495	0.315	0.597
	MD	-0.004	0.021	0.042	0.025	0.014	0.008
	MAD	0.125	0.098	0.138	0.115	0.133	0.109
	RMSD	0.175	0.140	0.185	0.155	0.160	0.159
b							
	Max D-	-1.514	-3.893	-0.660	-0.517	-1.750	-0.578
	Max D+	0.754	0.773	1.222	0.741	1.090	0.715
	MD	-0.083	-0.085	0.085	0.039	0.058	-0.011
	MAD	0.309	0.215	0.230	0.159	0.240	0.148
	RMSD	0.424	0.506	0.331	0.238	0.379	0.219
c							
	Max D-	-0.367	-0.163	-0.182	-0.291	-0.182	0.145
	Max D+	0.356	0.241	0.347	0.275	0.308	0.355
	MD	-0.004	0.003	0.025	0.009	0.016	-0.000
	MAD	0.062	0.040	0.073	0.053	0.073	0.045
	RMSD	0.094	0.061	0.102	0.077	0.099	0.076
ICC							
	Max IMD-	-0.081	-0.025	-0.098	-0.026	-0.088	-0.024
	Max IMD+	0.078	0.020	0.090	0.026	0.064	0.027
	MIMD	-0.000	0.000	0.000	-0.000	0.000	0.000
	RMMSD	0.040	0.020	0.041	0.023	0.044	0.024
	MMAD	0.029	0.014	0.029	0.016	0.033	0.016

[13] Criterion = same sample size calibration and scaling using paper-and-pencil data.

Appendix D

Verbal Items Flagged with Statistically Significant Item-Level Mode Effects

Section - Item	Logistic Regression Group	Logistic Regression Score x Group	Mantel-Haenszel	IRT
2 - 1				-
2 - 9				-
2 - 10				-
2 - 11		1		
2 - 14	1			
2 - 16		2		
2 - 17				-
2 - 18				+
2 - 19				+
2 - 20				+
2 - 21		1		+
2 - 23				-
2 - 25		1		
2 - 30				+
2 - 32				-
2 - 34	3	1		
2 - 36		1		
2 - 38		1		
5 - 1			1	-
5 - 2	3	1		-
5 - 5		1		+
5 - 6		1		
5 - 7		1		
5 - 8				-
5 - 10				+
5 - 11				-
5 - 12		3		
5 - 16				-
5 - 18	2	1		
5 - 24				-
5 - 27				+
5 - 28			2	-
5 - 29				+
5 - 30				-
5 - 31				-
5 - 33		1		
5 - 38		2		

Notes: Under Logistic Regression, "Group" refers to a significant ($p<.01$) group main effect, and "Score x Group" refers to a significant ($p<.01$) score-by-group interaction. Where both group and score-by-group interaction effects were found, only the interaction term is listed. The values "1," "2," and "3" refer to the groups being compared: "1" refers to CBT/T1GRE comparsion, "2" refers to the CBT/S-T1GRE comparison, and "3" refers to the S-T1GRE/T1GRE comparison. For the significant group effects, the first group listed in the pair always had a higher predicted probability of success on the item. Significant Mantel-Haenszel results always favored the CBT group. The IRT items flagged showed differences in b-values that were greater than one standard deviation of b-differences for all items; a positive sign means the CBT b-value was greater than the T1GRE b-value, and a negative sign means the CBT b-value was less than the T1GRE b-value.

Appendix D (continued)

Quantitative Items Flagged with Statistically Significant Item-Level Mode Effects

Section - Item	Logistic Regression Group	Logistic Regression Score x Group	Mantel-Haenszel	IRT
1 - 1				+
1 - 3			1	-
1 - 4		1		+
1 - 5		1		
1 - 6		1		
1 - 7		1		
1 - 10				-
1 - 12		1		
1 - 14				-
1 - 16	3			
1 - 20		1		+
1 - 22				+
1 - 23		1		
1 - 24	2			+
1 - 26			2,3	
4 - 3		1		
4 - 4				+
4 - 5			3	+
4 - 6	3		3	
4 - 7				+
4 - 11		1		+
4 - 16				+
4 - 17	1			
4 - 18				-
4 - 21				+

Notes: Under Logistic Regression, "Group" refers to a significant ($p<.01$) group main effect, and "Score x Group" refers to a significant ($p<.01$) score-by-group interaction. Where both group and score-by- group interaction effects were found, only the interaction term is listed. The values "1," "2," and "3" refer to the groups being compared: "1" refers to CBT/T1GRE comparsion, "2" refers to the CBT/S-T1GRE comparison, and "3" refers to the S-T1GRE/T1GRE comparison. For the significant group effects, the first group listed in the pair always had a higher predicted probability of success on the item, except for item 1-24, where the S-T1GRE group had a higher probability of success than the CBT group. Significant Mantel-Haenszel results always the first group listed in the pair, except for item 1-26, where the S-T1GRE group was favored over the CBT group. The IRT items flagged showed differences in b-values that were greater than one standard deviation of b-differences for all items; a positive sign means the CBT b-value was greater than the T1GRE b-value, and a negative sign means the CBT b-value was less than the T1GRE b-value.

Appendix D (continued)

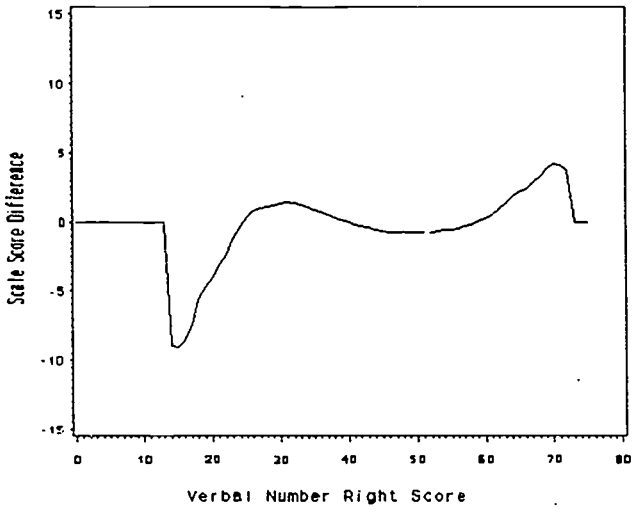
Analytical Items Flagged with Statistically Significant Item-Level Mode Effects

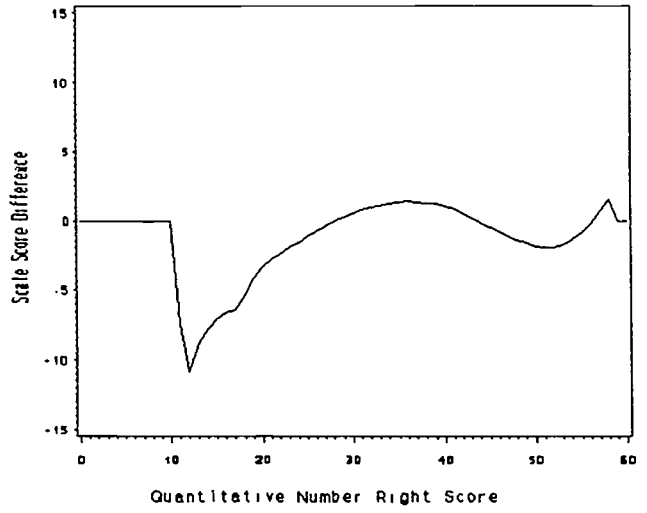
Section - Item	Logistic Regression Group	Logistic Regression Score x Group	Mantel-Haenszel	IRT
3 - 2		1		
3 - 3		1		-
3 - 4		1		
3 - 6			1,3	
3 - 7				-
3 - 8				+
3 - 9				+
3 - 11	1			+
3 - 15		1		
3 - 17	1	3		
3 - 18	1		2	
3 - 20	1		1	
3 - 22	1,3		1	
3 - 25	1			
6 - 2		1		
6 - 3	1		1	
6 - 15		1		
6 - 17	1			
6 - 23				+
6 - 25				+

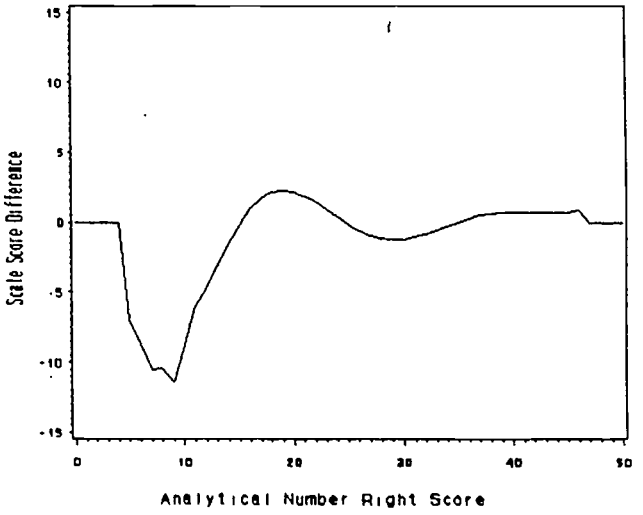
Notes: Under Logistic Regression, "Group" refers to a significant ($p<.01$) group main effect, and "Score x Group" refers to a significant ($p<.01$) score-by-group interaction. Where both group and score-by-group interaction effects were found, only the interaction term is listed. The values "1," "2," and "3" refer to the groups being compared: "1" refers to CBT/T1GRE comparsion, "2" refers to the CBT/S-T1GRE comparison, and "3" refers to the S-T1GRE/T1GRE comparison. For the significant group effects, the first group listed in the pair always had a higher predicted probability of success on the item. Significant Mantel-Haenszel results always favored the first group listed in the pair. The IRT items flagged showed differences in b-values that were greater than one standard deviation of b-differences for all items; a positive sign means the CBT b-value was greater than the T1GRE b-value, and a negative sign means the CBT b-value was less than the T1GRE b-value.

Appendix E

Plots of Differences in Equating Conversions (CBT-T1GRE)







Appendix F
CBT Questionnaire Results

A copy of the CBT questionnaire follows. The number next to each response is the percentage of the total group that marked that response.

Table F.1, which follows the copy of the questionnaire, shows the percentages of the total group and subgroups that selected each response to each item.

GRE® GRE® GRE® GRE® GRE®

NAME

LAST NAME (Family or Surname) | FIRST NAME (Given) | MI

REGISTRATION NUMBER

ENTER IF AVAILABLE

GRE QUESTIONNAIRE

Please circle the appropriate response to each question.

1. How often do you use a personal computer?

1 (1) Never before taking the test today (Skip to Question 4.)
18 (2) Rarely (only a few times in the last five years)
45 (3) Regular use (some time each week)
36 (4) Routine use (almost daily)

2. For what kinds of activities do you use a personal computer? Select as many as apply.

94 (1) Word processing
30 (2) Spreadsheets
28 (3) Database access
34 (4) Graphics
19 (5) Scientific computing
27 (6) Other software packages/applications
40 (7) Games
16 (8) Programming (list languages) _____ _____ _____
9 (9) Other (Explain.) _____

3. Do you own a computer, or is there a computer available to you where you live?

68 (1) Yes (Type.)_____
31 (2) No

4. Did you have any trouble reading the words and symbols on the computer screen?

10 (1) Yes
90 (2) No (Skip to question 6.)

5. If you answered yes to the previous question, which of the following were problems for you? Select as many as apply.

1 (1) The size of the print
1 (2) The characters were too close together to be legible.
2 (3) Too many words on some computer screens
3 (4) The contrast (the darkness of the letters against a bright background)
1 (5) The lighting in the room caused a glare on the screen.
4 (6) Graphs and/or diagrams
6 (7) Other (Explain.)_____

6. Was the computer monitor at a comfortable viewing angle?

(1) Yes, it was from the beginning.
(2) Yes, after I adjusted it.
4 (3) No

7. Did the computer screen flicker, making the questions difficult to read?

4 (1) Yes
96 (2) No

8. Did you experience eyestrain while taking this computer-administered test? Do you remember experiencing eye strain when you took the paper-and-pencil test in October 1991?

27 (1) None at either test session
30 (2) Some eyestrain at both test sessions
24 (3) More eyestrain at this test session
18 (4) More eyestrain at paper-and-pencil session

9. Are you wearing (Select one.)

22 (1) Contact lens
34 (2) Glasses
44 (3) No corrective lens

10. How often have you used a mouse on a personal computer?

15 (1) Never before taking this test
35 (2) A few times
29 (3) Regularly (some time each week)
20 (4) Routinely (almost daily)

11. How easy was it to use the mouse? Select all that apply.

77 (1) Very easy
15 (2) A little trouble at first
2 (3) Trouble for about half of the questions
0 (4) Trouble for most of the questions
5 (5) Trouble with the questions that required scrolling
8 (6) Trouble with navigating to the desired space on the computer screen

<u>The following questions ask about the Testing Tools.</u>

12. Did you use the "Time Option" to note the testing time remaining?

93 (1) Yes, the clock was always on.
5 (2) Sometimes the clock was on, sometimes it was off.
2 (3) No, I turned the clock off.

13. During the last five minutes of each section, did you find the "Time Display" in the upper left corner of the screen helpful or distracting?

65 (1) Helpful
4 (2) Distracting
23 (3) Helpful and distracting
6 (4) Ignored the time display

14. Were you confused by the concept of inactive (nonworking) testing tools (e.g. "Previous") being visible but light gray?

2 (1) Yes
94 (2) No
4 (3) No opinion or preference

15. Would you prefer to have the testing options (e.g. "Previous") disappear when inactive instead of turning light gray?

9 (1) Yes
48 (2) No
43 (C) No opinion or preference

16. For what purpose did you use "Help"? Mark all that apply.

78 (1) Did not use "Help."
18 (2) For test question directions
2 (3) For directions on How to Use the Testing Tools
1 (4) For directions on How to Scroll
3 (5) In what additional areas would you have wanted to use "Help"?

_____

_____

The next three questions ask about "Scrolling" (moving the text up and down).

17. How distracting was it to use the "Scrolling" option?

27 (1) Not distracting
26 (2) Little or no distraction
36 (3) Some distraction
11 (4) Much distraction

18. When you did scroll through passages or data, how easy was it to locate the information you were seeking?

28 (1) Very easy
45 (2) Somewhat easy
24 (3) Somewhat difficult
4 (4) Very difficult

19. How easy was it to use the computer to answer the questions that included a reading passage on one side of the screen and the question on the other side?

44 (1) Very easy
35 (2) Somewhat easy
17 (3) Somewhat difficult
4 (4) Very difficult

You were permitted to use "Review" only for the first six sections. You could not use "Review" during the last (seventh) section. The following questions ask you about the "Review" option.

20. How easy was it to use the "Review" option?

83 (1) Very easy
10 (2) Somewhat difficult
1 (3) Very difficult
7 (4) Did not use the "Review" option (Skip to question 26.)

21. Could the "Review" process be improved by adding a feature that would allow you to review a group of questions, e.g., all marked questions?

70 (1) Positive improvement
23 (2) No improvment
1 (3) Did not use "Review"

22. Did you wait until you finished the last question in the section before you used the "Review" screen?

61 (1) Yes
33 (2) No

23. For how many questions did you use the "Review" screen? Choose only one.

2 (1) None
59 (2) A few
25 (3) About one quarter
4 (4) One half
2 (5) Over one half
3 (6) Almost all

24. How did you use the "Review" screen? Select all that apply.

27 (1) To preview the test and to locate questions I wanted to answer first
58 (2) To determine which questions I had left unanswered
75 (3) To work on questions I had marked for review
10 (4) Other, (explain.) _____

25. Were the brief directions that preceded the question numbers on the "Review" screen helpful?

75 (1) Yes
17 (2) No

26. You were not permitted to use the "Review" option during the last (seventh) section. What was your reaction to this testing rule?

19 (1) Did not care
39 (2) Somewhat frustrating
35 (3) Very frustrating
42 (4) Comments _____

<u>The following questions ask you to compare your computer test-taking strategy with your paper-and-pencil test-taking strategy.</u>

27. Did you omit answers to

17 (1) More questions than on the paper-and-pencil test
47 (2) About the same number of questions as on the paper-and-pencil test
34 (3) Fewer questions than on the paper-and-pencil test

28. Did you omit questions but return to them later

18 (1) More often than on the paper-and-pencil test
39 (2) About as often as on the paper-and-pencil test
42 (3) Less often than on the paper-and-pencil test

29. Indicate if you used "Previous" or "Review" to return to unanswered and/or marked questions.

9 (1) Used "Previous."
29 (2) Used "Review."
57 (3) Used both.

30. Did you review questions you had answered

20 (1) More often than on the paper-and-pencil test
42 (2) About as often as on the paper-and-pencil test
(3) Less often than on the paper-and-pencil test

31. Did you change your original answers to questions

12 (1) More often than on the paper-and-pencil test
51 (2) About as often as on the paper-and-pencil test
37 (3) Less often than on the paper-and-pencil test

32. How do you think you would have done on a paper-and-pencil test with the same questions?

17 (1) Not as well on the paper-and-pencil test
59 (2) About the same
23 (3) Better on the paper-and-pencil test

33. If there were a computer test and a paper-and-pencil test with the same questions, which would you prefer to take?

60 (1) Computer
28 (2) Paper-and-pencil
10 (3) No preference
64 (4) Please explain your preference_____

_____

<u>The following questions deal with the test center environment.</u>

34. Was the GRE Test Center Instructions Sheet useful?

77 (1) Yes
20 (2) No

33 35. Please feel free to comment on the GRE Test Center Instructions Sheet.

_____

_____

_____

<u>Now that you have taken the computer test do you think there should be a mandatory break at the end of Section 3. If you took a break, skip to question 37.</u>

36. Do you think there should be a mandatory break?

25 (1) Yes
23 (2) No

37. Did you have enough space to use the mouse?

94 (1) Yes
6 (2) No

38. Did you have enough space to do scratch work?

84 (1) Yes
16 (2) No

39. Overall was the computer test center staff knowledgeable and able to answer questions?

67 (1) Yes, very knowledgeable
6 (2) Somewhat knowledgeable
1 (3) No, not knowledgeable
26 (4) Did not ask any questions

40. Was the lighting in the testing room adequate?

96 (1) Yes
3 (2) No (Describe.)_____

41. How would you rate the computer test-taking experience as compared with taking the paper-and-pencil test?

59 (1) Better than the paper-and-pencil test
22 (2) About the same as the paper-and-pencil test
13 (3) Worse than the paper-and-pencil test
79 (4) Comment briefly about your reactions, both positive and negative, to the computer-administered test. including positive and negative features.

_____

_____

_____

74 42. What kind of preparatory/background information could ETS/GRE provide prior to the computer test administration that would help the examinee adapt more easily to a test on a computer?

_____

_____

_____

65 43. Please feel free to comment on any aspect of this computer test.

_____

_____

_____

_____

Table F.1
CBT Questionnaire Percentages by Subgroup[14]

Item	T	F	M	A	B	H	W	Item	T	F	M	A	B	H	W
1.1	1	1	1	0	5	0	1	6.1	77	80	72	75	65	76	79
1.2	18	21	12	17	19	20	18	6.2	17	15	21	18	26	18	16
1.3	45	48	40	51	43	32	44	6.3	4	3	4	5	6	4	3
1.4	36	30	47	32	34	48	37	7.1	4	5	3	3	4	0	4
2.1	94	94	95	98	90	94	94	7.2	96	95	97	97	96	100	95
2.2	30	22	43	38	25	36	30	8.1	27	23	34	25	28	22	28
2.3	28	23	34	32	16	28	29	8.2	30	32	27	34	28	42	29
2.4	34	26	47	49	38	32	33	8.3	24	26	22	31	28	18	24
2.5	19	12	31	29	19	22	19	8.4	18	19	17	11	18	18	19
2.6	27	22	35	31	20	42	27	9.1	22	26	15	42	13	12	22
2.7	40	36	47	60	39	46	38	9.2	34	30	39	34	45	38	33
2.8	16	11	25	28	13	14	15	9.3	44	43	45	25	41	50	45
2.9	9	8	10	6	4	10	9	10.1	15	19	9	2	25	14	15
3.1	68	66	71	83	53	54	69	10.2	35	39	30	20	34	30	37
3.2	31	33	28	17	43	46	30	10.3	29	28	31	46	30	22	28
4.1	10	10	9	11	13	12	9	10.4	20	14	30	31	11	34	20
4.2	90	89	90	89	86	88	90	11.1	77	72	84	88	68	76	77
5.1	1	1	1	2	4	2	0	11.2	15	18	9	8	23	16	14
5.2	1	1	1	0	3	2	0	11.3	2	2	2	0	3	2	2
5.3	2	2	2	2	3	4	2	11.4	0	0	0	0	0	0	0
5.4	3	3	3	3	4	4	2	11.5	5	6	4	2	4	6	5
5.5	1	1	1	3	0	2	1	11.6	8	11	5	5	11	10	8
5.6	4	4	4	11	1	4	3	12.1	93	92	95	100	86	96	93
5.7	6	5	6	5	5	8	6	12.2	5	6	4	0	11	2	5

[14] Decimals omitted. T= Total Group (N=1,017); F= Female (N=617); M= Male (N=397); A= Asian (N=65); B= Black (N=80); H= Hispanic (N=50); W= White (N=782)

Table F.1
CBT Questionnaire Percentages by Subgroup[15] (continued)

Item	T	F	M	A	B	H	W	Item	T	F	M	A	B	H	W
12.3	2	2	1	0	3	2	2	19.1	44	42	46	48	34	40	45
13.1	65	63	68	68	68	62	65	19.2	35	34	36	35	41	44	33
13.2	4	6	3	5	5	4	4	19.3	17	18	15	15	16	12	17
13.3	23	24	23	26	18	30	24	19.4	4	5	3	2	5	4	4
13.4	6	6	6	2	8	4	7	20.1	83	81	86	82	73	86	83
14.1	2	2	2	2	6	6	1	20.2	10	11	8	14	16	6	9
14.2	94	94	93	91	88	92	95	20.3	1	0	1	0	0	0	1
14.3	4	4	5	8	5	2	4	20.4	7	7	6	5	9	8	7
15.1	9	10	8	12	15	14	8	21.1	70	70	70	69	71	76	69
15.2	48	45	53	43	49	42	49	21.2	23	22	24	26	18	16	23
15.3	43	45	39	45	34	42	43	21.3	1	1	1	0	1	0	1
16.1	78	79	76	77	70	72	79	22.1	61	64	56	54	58	54	63
16.2	18	16	21	20	24	22	16	22.2	33	28	39	42	33	40	31
16.3	2	2	2	2	3	4	2	23.1	2	2	2	0	3	2	2
16.4	1	1	0	0	0	2	1	23.2	59	56	65	51	55	62	61
16.5	3	3	3	3	4	0	3	23.3	25	28	19	35	24	22	24
17.1	27	26	29	31	33	30	26	23.4	4	4	4	3	3	6	3
17.2	26	21	32	22	30	30	25	23.5	2	2	2	5	1	2	1
17.3	36	40	29	29	26	38	37	23.6	3	2	4	2	5	0	3
17.4	11	12	9	18	9	2	11	24.1	27	24	31	38	21	28	26
18.1	28	25	33	26	36	22	27	24.2	58	57	60	74	55	50	58
18.2	45	44	46	45	36	54	45	24.3	75	73	78	80	58	76	77
18.3	24	27	18	25	20	22	24	24.4	10	9	11	11	11	10	9
18.4	4	4	3	5	5	2	3	25.1	75	75	74	66	83	76	75

[15] Decimals omitted. T= Total Group (N=1,017); F= Female (N=617); M= Male (N=397); A= Asian (N=65); B= Black (N=80); H= Hispanic (N=50); W= White (N=782)

Table F.1
CBT Questionnaire Percentages by Subgroup[16] (continued)

Item	T	F	M	A	B	H	W	Item	T	F	M	A	B	H	W
25.2	17	16	20	28	8	18	17	33.2	28	31	23	40	26	32	27
26.1	19	21	16	15	26	20	19	33.3	10	10	10	11	11	12	10
26.2	39	38	41	43	40	14	41	33.4	64	67	59	65	50	62	65
26.3	35	34	35	34	20	62	33	34.1	77	78	76	80	89	84	75
26.4	42	39	45	46	34	40	42	34.2	20	19	23	20	10	14	22
27.1	17	17	16	14	20	10	17	35	33	33	32	29	25	36	34
27.2	47	45	50	57	41	52	46	36.1	25	26	23	25	29	24	24
27.3	34	36	32	29	35	34	35	36.2	23	25	21	20	30	18	24
28.1	18	19	17	18	25	10	18	37.1	94	93	95	100	88	92	95
28.2	39	38	40	40	33	38	40	37.2	6	6	5	0	13	6	5
28.3	42	42	42	42	39	52	41	38.1	84	82	87	80	76	82	85
29.1	9	11	7	8	14	10	9	38.2	16	18	13	18	24	18	14
29.2	29	27	32	31	25	30	30	39.1	67	68	66	69	69	64	66
29.3	57	58	56	60	50	56	58	39.2	6	7	5	11	3	12	6
30.1	20	20	19	17	23	16	20	39.3	1	0	1	0	0	0	1
30.2	42	40	46	48	31	38	43	39.4	26	25	28	20	29	24	27
30.3	37	39	35	35	40	44	37	40.1	96	96	96	92	91	96	97
31.1	12	12	11	12	16	14	10	40.2	3	3	3	6	3	4	3
31.2	51	49	55	58	39	40	53	41.1	59	56	64	46	54	56	61
31.3	37	39	34	29	44	44	36	41.2	22	22	20	31	26	18	20
32.1	17	16	19	11	19	20	16	41.3	13	15	10	18	8	22	12
32.2	59	58	60	52	56	50	61	41.4	79	82	76	82	59	92	80
32.3	23	24	20	35	23	28	21	42	74	74	73	69	69	82	74
33.1	60	57	64	45	61	56	61	43	65	65	64	69	54	88	64

[16] Decimals omitted. T= Total Group (N=1,017); F= Female (N=617); M= Male (N=397); A= Asian (N=65); B= Black (N=80); H= Hispanic (N=50); W= White (N=782)

54020-07628 • Y43P.7 • 209090